A Validation Odyssey:
From big data to local intelligence
Validity of case-finding algorithms and of measures of quality of care for chronic diseases in Italian Administrative Databases

Rosa Gini

A Validation Odyssey:
From big data to local intelligence
Validity of case-finding algorithms and of measures of quality of care for chronic diseases in Italian Administrative Databases

Rosa Gini

The research described in this thesis was supported by the Agenzia regionale di sanità della Toscana, through the projects MATRICE and VALORE funded by the Italian Ministry of Health, and through the EMIF project funded by the Innovative Medicines Initiative.

ISBN: 978-94-6332-077-1

Lay-out: Ferdinand van Nispen, Citroenvlinder DTP & Vormgeving, *my*-thesis.nl
Printing: GVO drukkers & vormgevers, Ede, The Netherlands

A Validation Odyssey:
From big data to local intelligence
Validity of case-finding algorithms and of measures of quality of care for chronic diseases in Italian Administrative Databases

Een validatie-Odyssee: van 'big data' naar kennis
Validiteit van 'case-finding' algoritmes voor het meten van de kwaliteit van zorg voor chronische aandoeningen in Italiaanse Administratieve Databases.

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on

Friday 14 October 2016 at 11.30 hrs

by

Rosa Gini
born on Dolo, Italy

Erasmus University Rotterdam

Doctoral Committee:

Promotors:	Prof.dr. M.C.J.M. Sturkenboom Prof.dr. N.S. Klazinga
Other members:	Prof. F. Carinci Prof.dr. E.W. Steyerberg Prof. dr. A. Abu Hanna
Copromotor:	Dr. M.J. Schuemie

Dedicated to Eva Buiatti

TABLE OF CONTENTS

GENERAL INTRODUCTION

BACKGROUND AND FOCUS OF THIS THESIS

You take the blue pill, the story ends,
you wake up in your bed and believe whatever you want
to believe. You take the red pill, you stay in Wonderland,
and I show you how deep the rabbit hole goes.
Remember, all I'm offering is the truth, nothing more

Morpheus, in Matrix*, by The Wachowskis*

The era of big data opens new means to improve health and health care. Observational evidence from large populations can provide guidance to the choices of multiple stakeholders. Policymakers can understand how health care systems can be better organized, clinicians can explore more in detail all the treatment options, patients can put peculiarities of their own diseases at the centre of the clinical decisions, and citizens can obtain evidence to inform their political options. New methodologies need to be developed, and traditional tools and ways of thinking need to be renovated to adapt to the new perspective, to avoid drawing incorrect conclusions from studies based on this resource [Ray2011, Mooney 2015].

The methodological challenge relies not only in the observational nature of the studies that can be conducted on big data, but also in the heterogeneous characteristics of the data that is now available. We are interested in the latter.

Data is available from multiple sources such as hospitals and other health care settings, like primary care practices. Besides medical records, large administrative data sources are available, as well as disease registries, results from surveys etc. Record linkage between existing data sources is surrounded by data privacy concerns and possible in special circumstances only. According to national regulations and culture, record linkage and secondary use of data for the purposes of research is allowed to disparate types of organizations: university hospitals, public health institutions, scientific societies, private research companies [Trifiro2014].

In the situation of secondary use, variables are *derived* from existing data sources, by mean of data processing. This action mimics the traditional data collection process, but is in fact new: the 'true' variables should now be conceptualized as unobserved quantities, and the study variables entering the actual analysis as measurements, resulting from case-finding algorithms applied to the original data. The difference between the truth and the result of a case-finding algorithm can be assessed by validation studies. In the next subsections we review the existing methodology and applications of such studies, and describe the gap that we tried to address in this thesis.

Validation studies of case-finding algorithms

The objective of a validation study of a case-finding algorithm is estimating the *validity indices* of the algorithm. The indices quantify to which extent the study variable corresponds to the true variable. Imagine that we want to establish whether subjects have type 2 diabetes mellitus, but only have prescription information at our disposal to do so. We could adopt a case-finding algorithm that assigns a T2DM label if a record of prescribing metformin, a drug whose main indication is treatment of T2DM, is found. In this example, four subjects are classified according to whether they truly have type 2 diabetes mellitus (T2DM) and whether our algorithm would classify them as having T2DM: the first is the true variable, the second is the result of our case-finding algorithm (Table 1). In the example, Hellen is a *true positive* (she has T2DM, and has records of metformin prescriptions), Robert is a *false negative* (he has T2DM, but does not have an indication for treatment with metformin yet), Susan is a *false positive* (she does not have T2DM, but has polycystic ovary syndrome, a condition treated off-label with metformin) and Frank is a *true negative* (he does not have T2DM, and has no record of metformin prescription)

Table 1. A dataset with a true variable and a variable derived from an algorithm on a database of prescriptions.

Name	Has type 2 diabetes mellitus (true variable)	Was prescribed metformin (derived variable)
Hellen	Yes	Yes
Robert	Yes	No
Susan	No	Yes
Frank	No	No

Based on these concepts, the following validity indices are defined: *sensitivity* of this algorithm is the proportion of true positives among persons with T2DM; the *positive predictive value* is the proportion of true positives among persons who use metformin; *specificity* is the proportion of true negatives among persons without T2DM; *negative predictive value* is the proportion of true negatives among persons without prescription of metformin.
Results from a validation study of algorithms are used to adjust or, if this is not possible, to interpret the estimates obtained from analysing the dataset [Lanes2015]. Editorials and guidelines explicitly recommend the execution of this type of validation studies [GarciaRodriguez2010, Hernan2011, Benchimol2015].

Some methodological misconceptions hamper effective conduction and application of validation studies of case-finding algorithms.

Validation studies of case-finding algorithms versus diagnostic accuracy studies
The methodology for validation studies has been developed in the context of validation of *diagnostic tests*, that is, procedures collecting clinical parameters from patients [Whiting2003]. The diagnostic accuracy studies are typically devoted to estimate the likelihood that a person with specific values of the clinical parameters is indeed affected by the target health condition, and whether patients with the target condition are accurately captured by that same set of clinical parameters. Although the terminology and mathematical definition of validity indices is borrowed from diagnostic accuracy studies, the interpretation should be much different in studies on health care data that focus on identifying subjects with a certain disease from proxies that may indicate health status of a subject.
The parameters of sensitivity and specificity of a diagnostic test are rooted in human biology and essentially depend on the characteristics of the population where they are estimated [Greenland1996]. On the contrary, validity indices of a case-finding algorithm in a database depend on multiple characteristics of the system, for instance completeness of data collection, accuracy in coding habits, granularity of the coding system, organization of the health care system in the geographic area where the data is collected – besides characteristics of the population whose data is collected. All those factors are subject to change from one database to another, and over time [Quan2009, Reich2012, Herret2013, Morley2014, Rahimi2014, Lanes2015].

An effect of the difference between the two types of validation studies is the common misconception that predictive values depend on prevalence of the disease, while sensitivity and specificity do not [Greenland1996, Lanes2015b]. The rationale behind this statement is that mathematical equations link predictive values with prevalence, sensitivity and specificity [Altman1994, Altman1994b]. Indeed the equations apply, irrespectively of whether they refer to a diagnostic test or to an algorithm on a database, but in the latter case the four parameters *all depend on one another and on prevalence*.
The following thought experiment helps in understanding the difference. In an Italian factory with a workforce that is predominantly male, costs for the canteen are reduced, at the expenses of food quality. Many workers gain weight and new cases of T2DM are diagnosed, which *increases the prevalence* of the disease among the workforce. Some validity parameters of the algorithm mentioned in Table 1 ("Being prescribed metformin") are affected by this change. In Italy the first approach recommended after a diagnosis of T2DM is modification of life styles, without indication for pharmaceutical therapy [AMD-SID2014], therefore the new cases in the factory do not start a metformin treatment, and *the sensitivity of the algorithm is reduced*. On the contrary, *positive predictive value of the algorithm is not affected*: in the workforce there are no cases of the other common indication for metformin, polycystic ovary syndrome, and all patients utilizing metformin are cases of T2DM.

Validation studies of case-finding algorithms are difficult to generalize
Generalizing estimates of validity parameters of a case-finding algorithm outside of the environment where a validation study was executed, is questionable.

The construct of "validity of a diagnostic code" is recurrent in the literature of validation studies of case-finding algorithms [Cutrona2013, Valkhoff2014]. This construct is misleading: the same code recorded in primary care or during an emergency room visit, or in different countries adopting the same coding system, or even in the same country over time, may have completely different validity indices. Implying that a diagnostic code can be validated in itself may lead to inappropriate generalization of the validity indices.

In order to support effectively the interpretation of the results of a study, validation of study variables should be as close as possible to the actual dataset

that is used for the statistical analysis. Traditional validation studies, which imply manual assessment of samples of records, are time-consuming and expensive [Hernan2011].

An alternative solution is *exploiting big data to validate big data*: this methodological advancement is emerging in the recent literature.

Validation studies of case-finding algorithms exploiting existing data sources
In recent years, in several countries, many studies exploiting existing data sources for purposes of validating case-finding algorithms for diseases have been conducted [Lix2008, Lix2008b, Ferretti2009, Kahn2010, Amed2011, Gorina2011, Nosyk2013, Quantin2013, Bowker2015, John2016]. In each study at least two data sources were considered, with different data-generating mechanisms: administrative or claims databases, medical records, disease registries, survey responses, cohorts from epidemiologic studies. Two scenarios are common: first, when the same parameter can be estimated from different data sources at an **ecological level** on the same population; second, when the same variable can be estimated from different data sources at an **individual level** on a same set of individuals.

The first scenario has the advantage that no individual-level record linkage is requested between data sources belonging to different organizations. Ecologic parameters for comparison can be found in existing publications or national survey or census data. On the negative side, estimating validity parameters from this design is rarely possible.
The second scenario has the disadvantage that individual-level record linkage between different data sources, even for purposes of validation, often requires a complex legal permission procedure because of privacy considerations

Using existing data for validation purposes allows exploring large cohorts, often sampled from the general population, testing several algorithms for the same variable, identifying the main determinants of validity, and repeating the study over time.

Focus of this thesis

The main objective of this thesis is advancing the methodology of validation studies of case-finding algorithms that exploit diversity across available data, rather than collecting new data.

The case study that led us to this advancement was the assessment of the capacity of the Italian administrative database to capture cases of chronic disease to get estimates for the compliance with standards of care. Primary care medical records were the main comparative source. Part I of this thesis is focussed on this topic.

In Part II we exploited the results and extended the methodology of Part I to the context of multi-database, multi-national studies.

In the next section of this general introduction we describe in detail the research questions addressed in Part I and Part II.

INTRODUCTION TO PART I: VALIDATION OF VARIABLES DEFINING CHRONIC DISEASES AND COMPLIANCE WITH STANDARDS OF CARE IN ITALIAN ADMINISTRATIVE DATABASES

Italy has a universal, single-payer healthcare system. Chronic diseases impose an increasing burden on the Italian aging population, and are a major threat to sustainability of the healthcare system [OECD2015].
Administrative data are collected on a large set of services provided to the population, and are available for secondary analysis to health policy makers. Secondary use of Italian Administrative Databases (IAD) to detect patients with chronic conditions would allow surveillance, planning, monitoring of quality of healthcare, as well as assessment of impact of new organizational models on relevant health and quality outcomes.

Chronic diseases are normally diagnosed in a primary or secondary care setting in Italy. General practitioners (GP) have a gatekeeper role with respect to access to healthcare, but are paid a capitation fee, and don't feed administrative data. Specialist encounters contracted with the healthcare system are recorded, but no diagnostic code is included in the record. The sensitivity of the algorithms detecting diagnoses of chronic diseases from hospital discharge records was expected to be low. Algorithms using outpatient drug prescriptions or other sources of information were expected to be poor in positive predictive value, as these services can be provided for multiple indications. Moreover, it was relevant to understanding to which extent the validity of the algorithms depended on features of the local organization of the healthcare system, as this would have hampered the interpretation of comparisons.

The MATRICE Project, funded by the Italian Ministry of Health, was launched in 2011 by the Italian National Agency for Regional Healthcare Services (AGENAS), with the aim of defining methodologies and tools to best exploit administrative data for the purposes of monitoring quality of healthcare for patients with chronic diseases. Among other initiatives, the MATRICE Project funded this thesis.

Research questions

The main research question of Part 1 of the thesis was: what are the optimal algorithms to detect chronic diseases in the IAD, and what is the validity of estimates of compliance with standards of care?

We split this main question in 5 specific questions.

1. How do the prevalence estimates derived from finding cases of chronic diseases in IAD compare with estimates derived from other data sources?
2. What is the validity of algorithms detecting chronic diseases and their level of severity from medical records of the General Practitioners?
3. What are the optimal case-finding algorithms in IAD to find cases of chronic diseases?
4. How do estimates of compliance with standards of care derived from IAD compare with estimates derived from the Health Search (HSD), a database of medical records of the Italian College of General Practitioners?
5. How do measures of compliance with standards of care derived from IAD compare with measures derived from the medical records of the General Practitioners?

Figure 1 shows the research questions and study designs of the five chapters in a graphical manner.

Ecological level

In this thesis we used a dataset of administrative data collected during a previous project of AGENAS. Estimates of population prevalence of diabetes, ischaemic heart disease and chronic obstructive pulmonary disease could be obtained for five regions, and compared with the same parameters estimated from HSD and the national survey of the Italian Institute of Statistics (chapter 1). In chapter 4 the same dataset and HSD could be used to estimate compliance with standards of care for the same three conditions.

Individual level

The results of chapter 1 set the stage for chapter 2. We tested our assumption that querying automatically medical records of GPs led to a result they would have considered accurate themselves. We tested both presence of disease

Table 2. Summary of the studies in Part I of this thesis. IAD: Italian Administrative Databases.

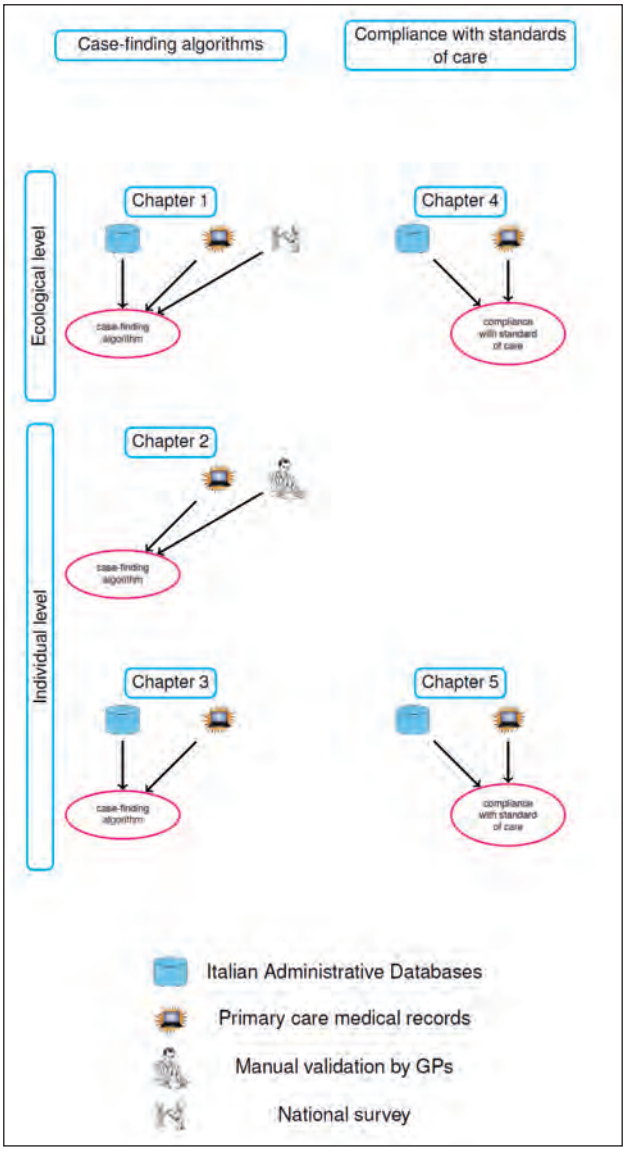
Variable to be validated	Study design	Data sources	Diseases	Setting	Year	Chapter
Case-finding algorithm	Ecological comparison between existing data	IAD, primary care medical records, national survey	Ischaemic heart disease, diabetes mellitus, heart failure, chronic obstructive pulmonary disease	Population samples from 5 Italian regions: Veneto, Emilia Romagna, Tuscany, Marche and Sicily	2009	1
Case-finding algorithm	Individual-level manual validation	Primary care medical records, manual assessment	Type 2 diabetes, hypertension, ischaemic heart disease and heart failure, with levels of severity	300 cases per disease, from 12 GPs across Italy	2014	2
Case-finding algorithm	Individual-level record-linkage between existing data	IAD, primary care medical records	Type 2 diabetes, hypertension, ischaemic heart disease and heart failure	25 clusters of subjects, 5 per each of the following regions: Lombardy, Veneto, Emilia Romagna, Tuscany, Puglia	2012	3
Compliance with standards of care	Ecological comparison between existing data	IAD, primary care medical records	Ischaemic heart disease, diabetes mellitus, heart failure	Population samples from 5 Italian regions: Veneto, Emilia Romagna, Tuscany, Marche and Sicily	2009	4
Compliance with standards of care	Individual-level record-linkage between existing data	IAD, primary care medical records	Type 2 diabetes, hypertension and ischaemic heart disease	25 clusters of subjects, 5 per each of the following regions: Lombardy, Veneto, Emilia Romagna, Tuscany, Puglia	2012	5

and their level of severity. This was the only traditional validation study in this thesis, and we used manual assessment by the GPs as a gold standard.

At the same time, we underwent the task of asking the National Privacy Authority permission to perform a record linkage, at an individual level, between IAD and medical records of 25 clusters of persons across the country, each counting almost 1,500 people and in charge to the same GP. When permission was granted we could perform the study as described in chapter 3: we were able to test a wide set of algorithms on IAD, considering GP medical records as a gold standard, for T2DM, hypertension and ischaemic heart disease (IHD). Moreover, in chapter 5, we used the same dataset to estimate compliance with standards of care for T2DM, hypertension and IHD in the same clusters, again comparing IAD with medical records. In this

case, medical records were not a gold standard. We could explore whether the cohorts detected by IAD were representative of the true cohorts, as long as compliance with standards of care was considered.

Figure 1. Graphic representation of the research questions and study designs of the chapters in Part I. On the left columns: studies of validity of case-finding algorithms for chronic diseases. On the right column: studies validating estimates of compliance with standards of care. On the top row: ecological studies. On the lower two rows: individual level study.



INTRODUCTION TO PART II: VALIDITY OF VARIABLES IN MULTI-DATABASE STUDIES

In 2004, after five year of widespread use, the anti-inflammatory drug rofecoxib was withdrawn from the market due to severe safety concerns. It was estimated that if a monitoring system had been in place querying the medical records of 100 million patients, the adverse cardiovascular effect would have been discovered in just few months. After this episode, networks of researchers with regular access to observational healthcare data sources have been created. Methods and procedures have been generated to execute studies in a distributed fashion, to take advantage both from size and from diversity of the populations they could merge [Trifiro2014]. Those advantages come at the price that the level of complexity in deriving study variables scales up, because all the characteristics that have an impact on validity may be different across sites.

This is especially true in Europe: diversity in local mechanisms of data collection is a consequence of the diversity among European countries in language, culture, political and health care organization. Notwithstanding, cross-border evidence is necessary to address common questions such as efficacy and safety of medical products and vaccines, or comparison of quality of health care.

Part II of this thesis was devoted to investigate the process adopted by some existing networks to derive the study variables in each site and address their validity, and to propose and test a novel methodology to streamline this process.

Research questions

The main research question of Part II was: how do networks of databases handle the process of generating study variables in the different sites, and how do they assess their validity?

We split the question in two parts.

6. How is the derivation of study variables organized in existing data networks in Europe and in the United States?
7. How can a network of diverse data sources in Europe streamline the process of data derivation?

In chapter 6 we introduced a conceptual framework representing the main data processing steps that a network needs to organize, and we used this framework to compare four case studies: the Italian MATRICE network, the European EU-ADR network, the OMOP and Mini-Sentinel networks from the United States. In particular we investigated how diversity in original data was documented and how the data derivation process was structured in the four networks.

In 2013 the Innovative Medicines Initiative, a joint undertaking between the European Union and the pharmaceutical industry association EFPIA, funded the European Medical Information Framework Project (EMIF), aimed to develop common technical and governance solutions and improve access and use of health data across Europe.

In chapter 7 we described a novel workflow introduced by EMIF to derive study variables, and to exploit diversity across data sources to obtain evidence on their validity. We applied the procedure to the case of type 2 diabetes mellitus in 8 European data sources. The Tuscan instance of Italian administrative databases and HSD were among the data sources involved in this study.

In the discussion at the end of this thesis the findings from this Italian validation Odyssey will be described and possibilities for generalization to other big-data settings that aim for the creation of local intelligence, will be reflected upon.

REFERENCES

[Altman1994] Altman DG, Bland JM. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. BMJ. 1994;308(6943):1552.

[Altman1994b] Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994;309(6947):102.

[AMD-SID2014] Italian Diabetologist Association, Italian Society for Diabetology. [Italian standards for treatment of diabetes mellitus] [In Italian]. http://www.standarditaliani.it/skin/www.standarditaliani.it/pdf/STANDARD_2014_May28.pdf

[Amed2011] Amed S, Vanderloo SE, Metzger D, Collet J -P, Reimer K, McCrea P, et al. Validation of diabetes case definitions using administrative claims data. Diabetic Medicine. 2011;28(4):424–7.

[Benchimol2011] Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. Journal of Clinical Epidemiology. 2011;64(8):821–9.

[Benchimol2015] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLOS Medicine. 2015;12(10):e1001885.

[Bowker2015] Bowker SL, Savu A, Lam NK, Johnson JA, Kaul P. Validation of administrative data case definitions for gestational diabetes mellitus. Diabet Med. 2015

[Carnahan2012] Carnahan RM. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data: summary of findings and suggestions for future research. Pharmacoepidemiology and Drug Safety. 2012;21:90–9.

[Cutrona2013] Cutrona SL, Toh S, Iyer A, Foy S, Daniel GW, Nair VP, et al. Validation of acute myocardial infarction in the Food and Drug Administration's Mini-Sentinel program. Pharmacoepidemiol Drug Saf. 2013;22(1):40–54.

[Ferretti2009] Ferretti S, Guzzinati S, Zambon P, Manneschi G, Crocetti E, Falcini F, et al. [Cancer incidence estimation by hospital discharge flow as compared with cancer registries data] [In Italian]. Epidemiol Prev.2009;33(4–5):147–53.

[GarciaRodriguez2010] García Rodríguez LA, Ruigómez A. Case validation in research using large databases. Br J Gen Pract. 1 2010;60(572):160–1.

[Gorina2011] Gorina Y, Kramarow EA. Identifying Chronic Conditions in Medicare Claims Data: Evaluating the Chronic Condition Data Warehouse Algorithm. Health Services Research. 2011;46(5):1610–27.

[Greenland1996] Greenland S. Basic methods for sensitivity analysis of biases. Int J Epidemiol. 1996;25(6):1107–16.

[Hernan2011] Hernan MA. With great data comes great responsibility: publishing comparative effectiveness research in EPIDEMIOLOGY. Epidemiology. 2011;22(3):290–1.

[Herret2013] Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, Staa T van, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. BMJ. 2013;346:f2350.

[John2016] John A, McGregor J, Fone D, Dunstan F, Cornish R, Lyons RA, et al. Case-finding for common mental disorders of anxiety and depression in primary care: an external validation of routinely collected data. BMC Med Inform Decis Mak. 2016

[Lanes2015] Lanes S, Brown JS, Haynes K, Pollack MF, Walker AM. Identifying health outcomes in healthcare databases. Pharmacoepidemiol Drug Saf. 2015;24(10):1009–16.

[Lanes2015b] Lanes S, Esposito D. Case ascertainment in electronic data bases: Where's Waldo? Community Meeting at Innovation in Medical Evidence Development and Surveillance (IMEDS). July 2015. http://imeds.reaganudall.org/sites/default/files/IMEDS_Outline_Dv8.pdf

[Kahn2010] Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. Br J Gen Pract. 2010;60(572):e128–36.

[Lix2008] Lix LM, Yogendran MS, Shaw SY, Burchill C, Metge C, Bond R. Population-based data sources for chronic disease surveillance. Chronic Dis Can. 2008;29(1):31–8.

[Lix2008b] Lix LM, Yogendran MS, Leslie WD, Shaw SY, Baumgartner R, Bowman C, et al. Using multiple data features improved the validity of osteoporosis case ascertainment from administrative databases. Journal of Clinical Epidemiology. 2008;61(12):1250–60.

[Mooney2015] Mooney SJ, Westreich DJ, El-Sayed AM. Commentary: epidemiology in the era of big data. Epidemiology. 2015;26(3):390–4.

[Morley2014] Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation. PLoS ONE. 2014;9(11):e110900.

[Nosyk2013] Nosyk B, Colley G, Yip B, Chan K, Heath K, Lima VD, et al. Application and Validation of Case-Finding Algorithms for Identifying Individuals with Human Immunodeficiency Virus from Administrative Data in British Columbia, Canada. Medeiros R, curatore. PLoS ONE. 2013;8(1):e54416.

[OECD2015] OECD. OECD Reviews of Health Care Quality. Italy 2014: Raising Standards. OECD Publishing; 2015.

[Quan2009] Quan H, Khan N, Hemmelgarn BR, Tu K, Chen G, Campbell N, et al. Validation of a Case Definition to Define Hypertension Using Administrative Data. Hypertension. 2009;54(6):1423–8.

[Quantin2013] Quantin C, Benzenine E, Velten M, Huet F, Farrington CP, Tubert-Bitter P. Self-controlled case series and misclassification bias induced by case selection from administrative hospital databases: application to febrile convulsions in pediatric vaccine pharmacoepidemiology. Am J Epidemiol. 2013;178(12):1731–9.

[Rahimi2014] Rahimi A, Liaw S-T, Taggart J, Ray P, Yu H. Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in Electronic Health Records. International Journal of Medical Informatics. 2014;83(10):768–78.

[Ray2011] Ray WA. Improving Automated Database Studies: Epidemiology. 2011;22(3):302–4.

[Reich2012] Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inform. 2012;45(4):689–96.

[Trifiro2014] Trifirò G, Coloma PM, Rijnbeek PR, Romio S, Mosseveld B, Weibel D, et al. Combining multiple healthcare databases for post-marketing drug and vaccine safety surveillance: why and how? J Intern Med. 2014

[Valkhoff2014] Valkhoff VE, Coloma PM, Masclee GMC, Gini R, Innocenti F, Lapi F, et al. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. Journal of Clinical Epidemiology. 2014;67(8):921–31.

[Whiting2003] Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Medical Research Methodology. 2003;3:25.

PART I

VALIDATION OF VARIABLES DEFINING CHRONIC DISEASES AND COMPLIANCE WITH STANDARDS OF CARE IN ITALIAN ADMINISTRATIVE DATABASES

CHAPTER 1

Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey

Rosa Gini [1,2], Paolo Francesconi [2], Giampiero Mazzaglia [3], Iacopo Cricelli [4], Alessandro Pasqua [3], Pietro Gallina [5], Salvatore Brugaletta [6], Daniele Donato [5], Andrea Donatini [7], Alessandro Marini [8], Carlo Zocchetti [9], Claudio Cricelli [3], Gianfranco Damiani [10], Mariadonata Bellentani [11], Miriam CJM Sturkenboom [2] and Martijn J Schuemie [2]

1. Agenzia regionale di sanità della Toscana, Osservatorio di epidemiologia; 50141 Florence, Italy
2. Department of Medical Informatics, Erasmus Medical Center; 3015 GJ Rotterdam, The Netherlands
3. Health Search, Italian College of General Practitioners and Primary Care; 50100 Florence, Italy
4. Genomedics; 50100 Florence, Italy
5. ULSS 16 Padova, Via Enrico Degli Scrovegni 14, 35131 Padua, Italy.
6. ASP 7 Ragusa, Piazza Igea 1, 97100 Ragusa, Italy.
7. Assessorato Politiche per la Salute, Viale Aldo Moro 21, 40127 Bologna, Italy.
8. Zona Territoriale Senigallia, Via Piero della Francesca 14, 60019 Senigallia (AN), Italy.
9. Regione Lombardia, Piazza Città di Lombardia 1, 20124 Milan, Italy.
10. Università Cattolica del Sacro Cuore, Largo Francesco Vito 1, 00198 Rome, Italy.
11. Agenzia Nazionale per i Servizi Sanitari Regionali; 00100 Rome, Italy

Published in
Gini et al. BMC Public Health 2013, 13:15
http://www.biomedcentral.com/1471-2458/13/15

ABSTRACT

Background

Administrative databases are widely available and have been extensively used to provide estimates of chronic disease prevalence for the purpose of surveillance of both geographical and temporal trends. There are, however, other sources of data available, such as medical records from primary care and national surveys. In this paper we compare disease prevalence estimates obtained from these three different data sources.

Methods

Data from general practitioners (GP) and administrative transactions for health services were collected from five Italian regions (Veneto, Emilia Romagna, Tuscany, Marche and Sicily) belonging to all the three macroareas of the country (North, Center, South). Crude prevalence estimates were calculated by data source and region for diabetes, ischaemic heart disease, heart failure and chronic obstructive pulmonary disease (COPD). For diabetes and COPD, prevalence estimates were also obtained from a national health survey. When necessary, estimates were adjusted for completeness of data ascertainment.

Results

Crude prevalence estimates of diabetes in administrative databases (range: from 4.8% to 7.1%) were lower than corresponding GP (6.2%-8.5%) and survey-based estimates (5.1%-7.5%). Geographical trends were similar in the three sources and estimates based on treatment were the same, while estimates adjusted for completeness of ascertainment (6.1%-8.8%) were slightly higher. For ischaemic heart disease administrative and GP data sources were fairly consistent, with prevalence ranging from 3.7% to 4.7% and from 3.3% to 4.9%, respectively. In the case of heart failure administrative estimates were consistently higher than GPs' estimates in all five regions, the highest difference being 1.4% vs 1.1%. For COPD the estimates from administrative data, ranging from 3.1% to 5.2%, fell into the confidence interval of the Survey estimates in four regions, but failed to detect the higher prevalence in the most Southern region (4.0% in administrative data vs 6.8% in survey data). The prevalence estimates for COPD from GP data were consistently higher than the corresponding estimates from the other two sources.

Conclusion
This study supports the use of data from Italian administrative databases to estimate geographic differences in population prevalence of ischaemic heart disease, treated diabetes, diabetes mellitus and heart failure. The algorithm for COPD used in this study requires further refinement.

BACKGROUND

Administrative healthcare data are collected by privately owned health maintenance organisations or governmentrun institutions for managerial reasons. Due to differences in healthcare systems, the content of the administrative data may vary from country to country. They may contain records collected at hospital discharge, during encounters with the general practitioner (GP) or specialist, at drug prescription or dispensation, or upon request for, or conduct of, a diagnostic analysis or procedure. The content also depends on the choices of the organisation: data may or may not contain diagnosis codes; drug prescriptions may or may not contain indication of use, data from laboratories may or may not contain the actual result.

Secondary use of administrative healthcare data has been increasing over the years, including the provision of prevalence estimates for chronic diseases [1], such as diabetes mellitus, chronic obstructive pulmonary disease (COPD), hypertension, ischaemic heart disease, cerebrovascular disease, and depression. Case finding and ascertainment algorithms are tailored to the structure and type of information that is captured in the specific administrative database. Sensitivity and specificity of such algorithms are conditioned on distinguishing features such as presence of drug dispensing as well as sources for diagnostic codes. The Canadian [2], Swedish [3] and Medicare [4] administrative databases, for example, contain diagnosis codes from hospitalization episodes as well as from outpatient care, hence enriching the data for estimation of chronic disease prevalence. In settings where outpatient care diagnoses are not available, other solutions have been explored. For instance, in Luxembourg and France, where only drug prescriptions are available, diabetes could be identified in treated patients by analysing the volume of prescriptions for anti-diabetic drugs [5].

Observational studies based on administrative databases need careful validation of the algorithms they rely on in order to provide sound epidemiologic research [6]. Validation of chronic disease case ascertainment algorithms has been performed either through direct [7] or indirect [8] clinical assessment or through individual record linkage with other electronic data sources, such as disease registries [9] or health surveys [4,10]. When individual record linkage with non-administrative data sources was not feasible, the performance of

the algorithms has been inferred though external comparison with prevalence estimates obtained from health surveys [5,11]. Italy has a tax-based, universal coverage national health system organised in three levels: national; regional (21 regions); and local (on average 10 local health units per region) [12]. Administrative data on healthcare reimbursed by the system, such as inpatient care and drug dispensations, are routinely collected by local health units and, in some regions, sent to the regional level. Transmission to the national level is obligatory, and a common data model for data transmission is mandated by law on a national level. Before data are sent to the national level, however, unique personal identifiers are removed, hence record-linkage cannot be performed outside a single region. The Italian administrative databases therefore form a virtual national information system, with homogeneous data collected at the local level. Actual databases allowing record-linkage only exist up to the regional level. Data on diagnosis collected in outpatient settings are not part of the Italian administrative databases, therefore algorithms for case ascertainment developed in other countries that make use of this information cannot be applied to the Italian situation. Several studies have investigated the comparison between chronic disease prevalence estimates from Italian administrative data and other Italian data sources [11,13]. However studies to date were only performed within local or regional databases. Capture–recapture technique, a more sophisticated analysis aimed at estimating a suspected underascertainment when more than two lists of cases are available [14,15], was applied as well in estimating diabetes prevalence from administrative databases, in specific geographic areas of the country [16,17]. In light of the strong difference in health and healthcare quality across Italy [18] it is relevant to understand whether administrative databases can support chronic disease prevalence surveillance in different areas of the country.

In 2010 the Italian national project VALORE was launched aimed at assessing quality of care for chronic diseases in five different Italian regions, based on secondary use of data from administrative databases. In this study we describe prevalence estimates for diabetes mellitus, heart failure, ischaemic heart disease and COPD from these data and compare the estimates with prevalence estimates obtained from a national GP electronic medical record database and, where possible, from a national health survey.

METHODS

Setting

The five regions which contributed data to the VALORE study were: Veneto (A, Northern Italy), Emilia Romagna (B, Northern Italy), Tuscany (C, Central Italy), Marche (D, Central Italy) and Sicily (E, Southern Italy). The following data files were used in the VALORE project:

- **Hospital discharge records** with one main and five secondary diagnoses coded using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9CM);
- **Drug dispensing records** coded using Anatomic Therapeutic Chemical (ATC) codes for drug classification; the ATC system is the drug classification system adopted by the World Health Organization [19];
- **Disease-specific exemptions** from copayment to health care coded using ICD9CM;
- **Inhabitant Registry** (IR) with demographic information (birthyear, gender) and identifier of the GP in charge.

In each region, record-linkage within and between data files was done deterministically with a unique coded personal identifier. Region B could not provide the file of exemptions from copayment. For organizational reasons, the regions participating in the VALORE project did not provide administrative data of the whole regional population, but only of specific geographical subareas. In each region raw data were extracted from the local data files and sent via File Transfer Protocol (FTP) to a single data management center, after anonymization of the coded personal identifier. A standardized automated routine was developed in Stata 9.2 to apply the case ascertainment algorithm and to calculate the prevalence estimates. Each regional sample consisted of all inhabitants registered in the selected geographical subareas and alive at the index date (January 1st 2009).

Case ascertainment

The case finding and ascertainment algorithms that were used to detect the specific diseases are shown in Table 1. Regional administrative databases link Hospital discharge records (HOSP), Drug dispensation records (DRUG), and Disease-specific exemptions (EXE) from 2003 to 2008, and a patient was classified as having the selected disease if at least one of the corresponding conditions listed in Table 1 were met, i.e. condition 1 OR condition 2 OR

condition 3. For 93% of the population, the full six years of follow-up data were available and were included in the analysis.
Diverse algorithms for diabetes, COPD and ischaemic heart disease case ascertainment from Italian administrative databases have been previously described in theliterature. Those published in Simonato et al. [20] were the result of a workgoup involving two Italian scientific associations of epidemiologists and biostatisticians, and were therefore adopted. However, to deal with a previously reported issue of lack of sensitivity, the algorithm for COPD was enriched with drug dispensing data [21]. In addition we also calculated a prevalence estimate of diabetes based on anti-diabetic treatment alone. The heart failure algorithm was defined in the VALORE project.

Comparison data
The Health Search Database (HSD) collects electronic medical record data from a network of Italian GPs who are members of the Italian College of General Practitioners [22]. The GPs participating in HSD all use the same information software, in which they record demographic information, visits and referrals, diagnoses (both in free text and ICD9CM codes), drug prescriptions and clinical information. For this study, data from 199 GPs practicing in one of the five regions of the VALORE project were used. The study population comprised patients aged 16- 95 who had been enrolled for at least two years and were alive on 1st January 2009. Prevalence estimates were calculated based on the number of patients enrolled with the GPs at the index date (January 1st 2009) as denominator. The numerator represented all cases with specific diseases as ascertained through a query in the PROBLEM field of the clinical database, where diagnoses are coded. The diagnosis codes are shown in Table 1.
The Italian National Health Survey is conducted every five years by the National Institute of Statistics (ISTAT). In addition, there is a yearly survey that captures relevant health-related issues, and in particular diabetes and COPD. The 2008 survey in the five regions of the VALORE study comprised 11,656 people aged 16 years and above. The survey sample was extracted according to a two-stage weighted cluster sampling design (first level: municipalities; second level: families). Answers to two questions were used for this study: Are you affected by one or more of the following chronic diseases? Diabetes (Yes/No) Chronic bronchitis, emphisema, repiratory failure (Yes/No). Questions about heart failure and ischaemic heart disease were not asked.

Data analysis

Administrative data provided estimates of prevalence in three different ways: (1) analysis of distribution of prevalence per GP practices; (2) pooled analysis; (3) capturerecapture analysis. While (2) and (3) provided estimates that could be compared with the regional estimates from the other two data sources, it must be noted that as the population sample from each regional population is not random, but rather geographically restricted, it was expected that the comparison was biased. The rationale for analysis (1) was therefore the following: as both administrative and GP data could be aggregated per practice and as practices could be considered to be (non randomly) sampled from the same population of regional practices, if the two measurement techniques (administrative versus GP data) did measure in fact the same population parameter it was expected that the pairs of regional distributions overlapped and distributions within the same region were more similar to each other than distributions within the same data source.

All the analyses refer to the population aged≥16 years, both male and females, although in the GP database ages above 95 were truncated. Sex and age distribution of each regional sample were computed. The percentage of the regional population covered by the sample was estimated by dividing its number by the estimates of the regional population according to the National Institute of Statistics [23]. Prevalence was estimated as the total number of existing cases divided by the number of subjects in the sample. Every adult Italian inhabitant is entitled to choose a GP, and GPs may accept a maximum of 1,500 patients [12]. For each GP the population registered with that GP at the index date was calculated and used as denominator for the prevalence estimates. Median and interquartile (IQ) range of this distribution was computed in each regional sample. To avoid spurious results the disease prevalence per GP practice was computed only for those practices who had at least 300 people enrolled and at least 4 patients with the disease.

In the Health Search Database the same prevalence measures were estimated, the sample being the number of inhabitants in charge of the GPs of HSD at the index date. From the National Health Survey the variable of the first-level sampling design (municipalities) was not available, hence simple weighted analysis was performed, with probability weight attributed to each individual. Finally, to ascertain the degree of completeness in capturing diabetes cases from administrative data, a capturerecapture analysis was performed. Log-linear

models were estimated by sequentially incorporating pair-wise dependency between sources, andmodel selection was based on the Akaike information criterion (AIC) criterion. This was not done for region B, where only two sources of data were available, since independence between two data sources could not be assumed [15].

All analyses were performed with Stata 9.2.

Table 1. Case ascertainment algorithms for diabetes, ischaemic heart disease, heart failure and COPD

Disease	Administrative data			GP data
	HOSP (ICD9CM) +	DRUG (ATC)++	EXE (ICD9CM)	PROBLEM (ICD9CM)
Diabetes mellitus	250*	A10	250	250*
Treated diabetes		A10		250* AND A10 ++++
Ischaemic heart disease	410-*414*	C01DA	414	410-*414*
Heart failure	428*, 40201, 40211, 40291, 40401, 40403, 40411, 40413, 40491, 40493	-	428	428*, 40201, 40211, 40291, 40401, 40403, 40411, 40413, 40491, 40493
COPD	490*-492*, 494*, 496*	R03 +++	-	490*-492*, 494*, 496*

+ Either in main or in one of the secondary diagnoses
++ At least two dispensations in different dates in a single year
+++ A specific algorithm involving number, heterogeneity of ATC codes and time span of dispensations is used, see [Anecchino2007]
++++ Patients having at least 2 prescriptions in one of the previous 2 years

Ethical approval

No identifiable human data were used for this study. The dataset used in the study is not openly available. Permission to use non-identifiable, individual data extracted from administrative databases for the VALORE project was granted by ULSS 16 Padova, ASP 7 Ragusa, Assessorato Politiche per la Salute Emilia Romagna, Zona Territoriale Senigallia, which are responsible for the use of the data of the corresponding populations. Agenzia regionale di sanit`a della Toscana is enabled by a regional law to use Tuscan data for research purposes. Approval for use of encrypted and aggregated data from the HSD was also obtained from the Italian College of General Practitioners. Data from the National Health Survey are openly available from ISTAT.

RESULTS

The subpopulations whose data were collected covered a percentage of the total population of the regions, as shown in Table 2. The age distribution of the three population samples in the five regions is shown in Figure 1. There were some differences in age distribution between the populations from the different sources (see Figure 1). The prevalence estimates in the five regions from the three sources are shown in Figure 2 and Table 3. Administrative data underestimated the prevalence of diabetes as compared to both GP and survey estimates across most regions although differences were often barely or non significant and the increasing North-South trend could be consistently observed in the three sources. Adjustement for underascertainment led to higher estimates with respect to both GP and Survey figures, except in one region. The width of the interquartile range (IQ) of the practice-level estimates was higher in GP data than in administrative databases. When prevalence of diabetes was estimated based only on diabetes treatment, the prevalence

Figure 1. Age distribution in each region from each data source. Age distribution in each region of the sample extracted from administrative databases (Admin), of the sample extracted from clinical data collected by GPs participating to the Health Search Database (GP) and of the sample participating to the National Health Survey (Surv).

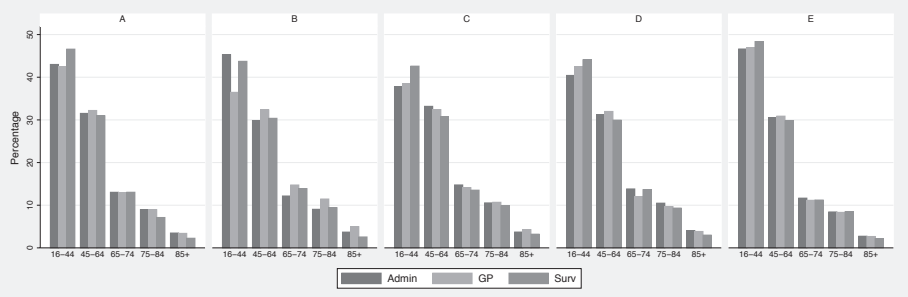
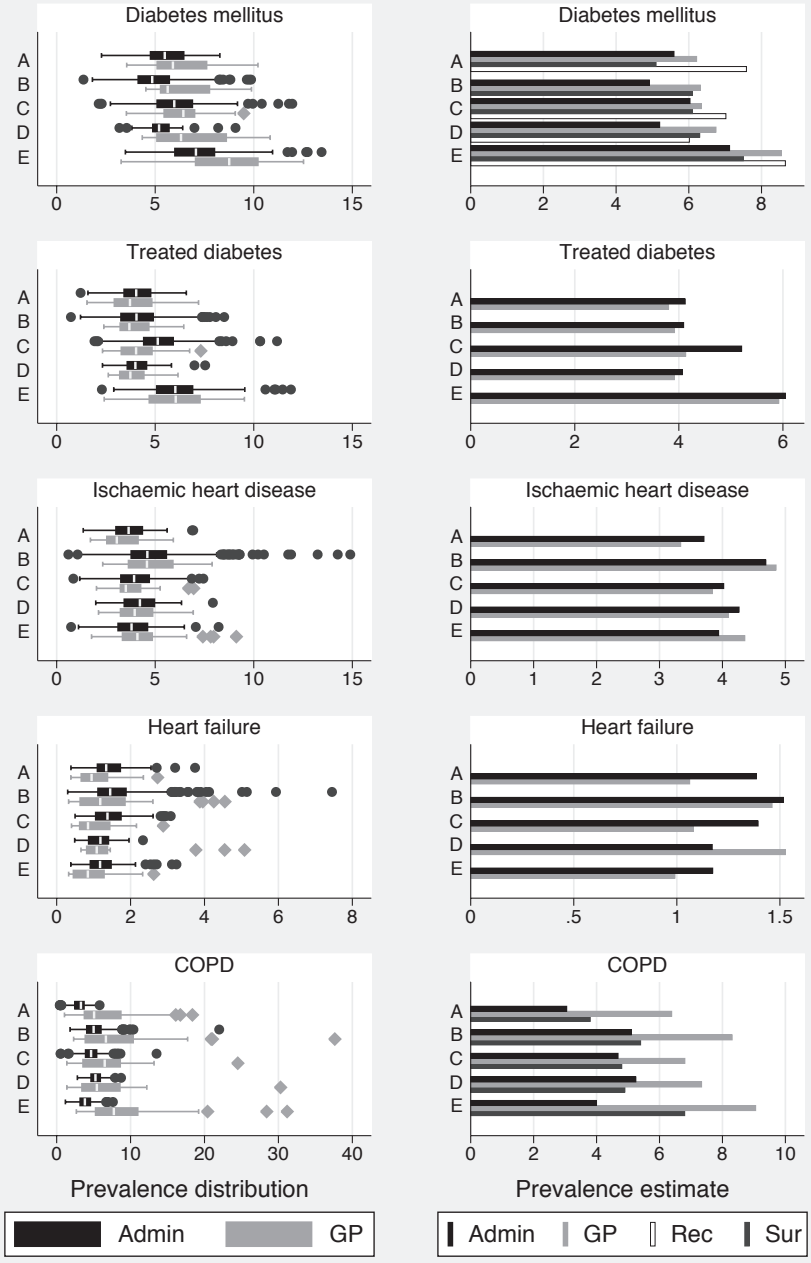


Figure 2. Prevalence estimates for diabetes mellitus, treated diabetes, ischaemic heart disease, heart failure and COPD from each data source. Crude prevalence estimates for diabetes mellitus, treated diabetes, ischaemic heart disease, heart failure and COPD in 5 Italian regions, according to administrative data (Admin) and clinical GP data (GP) and, for diabetes and COPD only, the National Health Survey (Sur). For diabetes mellitus estimates from administrative data adjusted for ascertainment are also presented (Rec). On the left column prevalence is represented by box plots of the distribution of the disease prevalence in GP practices: the central line is the median value, the box covers the interquartile range, while wiskers range from a minimum to a maximum value except for some observations which are detected as outliers and are representes as single dots or diamonds; comparison is only between GP and Admin data sources. On the right column prevalence is represented as global estimate. Date: 1 January 2009 Population: male and females, aged 16+.


Diabetes mellitus
Diabetes mellitus
Treated diabetes
Treated diabetes
Ischaemic heart disease
Ischaemic heart disease
Heart failure
Heart failure
COPD
COPD
A
B
C
D
E
Prevalence distribution
Prevalence estimate
Admin
GP
Rec
Sur

estimates obtained from GP data were fairly consistent with those obtained from administrative data in all regions; the width of the IQ range of the practice estimates was similar between GP data and administrative data in all regions. The prevalence of ischaemic heart disease, as estimated using administrative data, was similar to that estimated using GP data, prevalence ranging from 3.3% (region A, source GP) to 4.9% (region B, source GP). The width of the IQ range was similar between GP data and administrative data in all regions.

The prevalence estimates for heart failure were lower in GP data than in administrative data in three regions, with the highest difference (1.1% vs 1.4%) observed in both regions A and C, where significance was observed as well. According to age-specific prevalence estimates shown in Figure 3, the difference in estimates were increasing with age. The width of IQ range of the practice estimates was higher for GP data in three regions. The prevalence estimates from administrative data for COPD in regions A to D fell within the confidence interval of the survey estimates and ranged from 3.8% (region A) to 4.9% (region D), but failed to detect the higher prevalence in Region E (4.0% in administrative data vs 6.8% in survey data). The estimates from GP data were consistently higher, ranging from 6.4% (region A) to 9.1% (region E), and detected the same increased prevalence in region E. The width of IQ range of the practice estimates was much higher for GP data. In all diseases the width of the IQ range of the practice estimates were fairly consistent across regions in administrative data. Mean and medians were pretty close in both sources and all regions, and, except for outliers, distributions were rather symmetric, although in GP data data the distribution was slightly skewed to the right for diabetes and COPD.

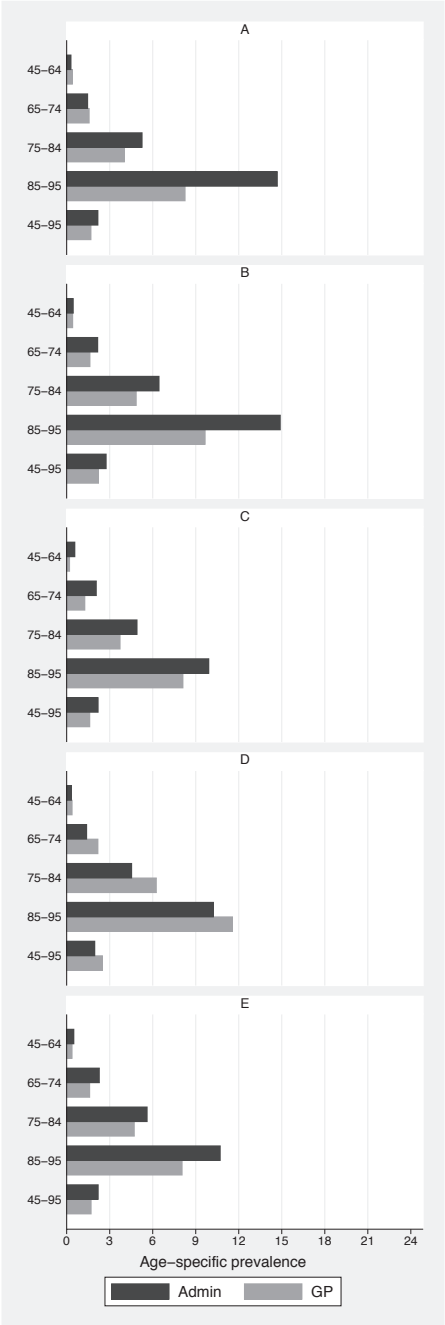
Table 2. Characteristics of the subpopulations of each region covered, respectively, by administrative or GP data. Data on general population from the Italian National Institute of Statistics. Analysis is restricted to inhabitants aged 16+

Region	Population aged 16+ (millions)	Administrative data coverage			GP data coverage		
		N GPs	N patients	% pop	N GPs	N patients	% pop
A	4.2	140	167,805	4.0	51	70,301	1.7
B	3.7	625	840,546	22.5	41	60,590	1.6
C	3.2	511	498,084	15.5	29	36,908	1.1
D	1.3	57	63,125	4.7	18	24,912	1.8
E	4.2	231	264,902	6.3	60	84,483	2.0

Table 3. Crude prevalence estimates for diabetes mellitus, treated diabetes, ischaemic heart disease, heart failure and COPD in 5 Italian regions, according to administrative data (Admin) and clinical GP data (GP). For diabetes and COPD only: crude prevalence estimated from the National Health Survey (Surv). For diabetes only: prevalence estimates from administrative data adjusted for estimated completeness of ascertainment (Admin-Recap). Prevalence estimated both as global percentage with 95% confidence interval and, for Admin and GP only, as median with interquantile range of the distribution of prevalence in GP practices. Date: 1 January 2009. Population: male and females, aged 16-95 in GP and 16+ in the other sources.

		A		B		C		D		E	
Disease	Data source	Mean (95% CI)	Median (IQ range)	Mean (95% CI)	Median (IQ range)	Mean (95% CI)	Median (IQ range)	Mean (95% CI)	Median (IQ range)	Mean (95% CI)	Median (IQ range)
Diabetes mellitus	Admin	5.6 (5.5-5.7)	5.5 (4.8-6.4)	4.9 (4.9-5.0)	4.8 (4.1-5.7)	6.0 (6.0-6.1)	6.0 (5.1-6.9)	5.2 (5.0-5.4)	5.2 (4.9-5.7)	7.1 (7.0-7.2)	7.1 (6.0-8.0)
	GP	6.2 (6.0-6.4)	5.9 (5.1-7.6)	6.3 (6.1-6.5)	5.6 (5.3-7.7)	6.3 (6.1-6.6)	6.4 (5.5-7.0)	6.7 (6.4-7.1)	6.3 (5.1-8.6)	8.5 (8.4-8.7)	8.8 (7.0-10.2)
	Survey	5.1 (4.2-6.0)		6.1 (5.1-7.1)		6.1 (5.1-7.1)		6.3 (5.2-7.4)		7.5 (6.4-8.5)	
	Admin-Recap	7.6 (7.1-8.2)				7.0 (6.8-7.3)		6.0 (5.8-6.4)		8.7 (8.4-9.0)	
Treated diabetes	Admin	4.1 (4.0-4.2)	4.0 (3.4-4.8)	4.1 (4.0-4.1)	4.0 (3.3-4.9)	5.2 (5.1-5.3)	5.1 (4.4-5.9)	4.1 (3.9-4.2)	4.0 (3.6-4.6)	6.0 (6.0-6.1)	6.0 (5.1-6.9)
	GP	3.8 (3.7-3.9)	3.7 (2.9-4.8)	3.9 (3.8-4.1)	3.7 (3.2-4.7)	4.1 (3.9-4.3)	4.0 (3.3-4.8)	3.9 (3.7-4.2)	3.7 (3.2-4.4)	5.9 (5.8-6.1)	6.0 (4.7-7.3)
Ischaemic heart disease	Admin	3.7 (3.6-3.8)	3.6 (3.0-4.3)	4.7 (4.6-4.7)	4.6 (3.8-5.6)	4.0 (4.0-4.1)	3.9 (3.3-4.7)	4.3 (4.1-4.4)	4.2 (3.4-5.0)	3.9 (3.9-4.0)	3.8 (3.1-4.6)
	GP	3.3 (3.2-3.5)	3.1 (2.5-4.1)	4.9 (4.7-5.0)	4.6 (3.6-5.9)	3.8 (3.6-4.0)	3.5 (3.2-4.3)	4.1 (3.8-4.3)	4.0 (3.2-4.9)	4.4 (4.2-4.5)	4.1 (3.3-4.8)
Heart failure	Admin	1.4 (1.3-1.4)	1.3 (1.1-1.7)	1.5 (1.5-1.5)	1.4 (1.1-1.9)	1.4 (1.4-1.4)	1.4 (1.1-1.7)	1.2 (1.1-1.3)	1.2 (0.9-1.4)	1.2 (1.1-1.2)	1.2 (0.9-1.5)
	GP	1.1 (1.0-1.1)	0.9 (0.7-1.4)	1.5 (1.4-1.6)	1.2 (0.6-1.8)	1.1 (1.0-1.2)	0.8 (0.6-1.4)	1.5 (1.4-1.7)	1.1 (0.8-1.4)	1.0 (0.9-1.1)	0.8 (0.5-1.3)
COPD	Admin	3.1 (3.0-3.1)	3.3 (2.5-3.7)	5.1 (5.1-5.2)	5.0 (4.1-6.0)	4.7 (4.6-4.7)	4.7 (3.9-5.4)	5.2 (5.1-5.4)	5.2 (4.7-5.9)	4.0 (3.9-4.1)	3.8 (3.2-4.6)
	GP	6.4 (6.2-6.6)	5.1 (3.8-8.7)	8.3 (8.1-8.5)	6.7 (3.9-10.3)	6.8 (6.5-7.1)	6.5 (3.6-8.6)	7.3 (7.0-7.7)	5.4 (3.4-8.6)	9.1 (8.9-9.2)	7.7 (5.3-11.0)
	Survey	3.8 (3.0-4.6)		5.4 (4.4-6.4)		4.8 (3.9-5.7)		4.9 (3.8-5.9)		6.8 (5.8-7.8)	

Figure 3. Age-specific prevalence of heart failure. Age-specific prevalence of heart failure in 5 Italian regions, according to administrative data (Admin) and clinical GP data (GP). Date: 1 January 2009 Population: male and females, aged 16+



DISCUSSION

Overall, differences in prevalence estimates among the different sources in a region were lower than the differences between regions, and differences observed among regions were similar across data sources. The fact that independent sources of data showed consistent values across different regions supports the claim that they correspond to actual populationmeasures. In case systematic differences were observed, they could be interpreted as being due to differences in data collection and associated to demographic and disease characteristics. This provides evidence that administrative data actually measure a population phenomenon that can be interpreted and supports the use of administrative data for surveillance of geographical trends of the diseases in study, with the possible exception of COPD.

In the case of diabetes mellitus, the observed concordance between estimates from GP data and from survey data confirms previous reports [24,25]. Estimates from GP data were systematically higher than estimates from administrative data. According to reports from other countries [26], the difference is likely to be due to the proportion of patients who, although being diagnosed with diabetes to the knowledge of their GPs, have mild or well-controlled disease and thus have never had either a hospital admission or a prescription for antidiabetic drugs, and have not received an exemption from copayment of diabetes-related healthcare, therefore escaping the algorithm for administrative databases. Indeed, when the subset of patients undergoing therapy with antidiabetics in the previous two years were extracted from both administrative and GP data sources, the pairs of prevalence estimates almost coincided in all of the regions, with one exception. Estimates adjusted for completeness of ascertainment, on the other hand, provided slightly higher estimates, a finding consistent with a previous study with similar data in another Italian area [16].

Ischaemic heart disease being congruently estimated by administrative and GP data in all of the regions is an unexpected finding. Angina, a less severe form of the disease, does not lead per se to a hospital admission, and few cases (less than 5%) are detected by the registry of exemptions from copayment. As around 30% of cases are detected only by dispensings of nitrates (data not shown), we observe that nitrates therapy is probably specific in detecting cohorts bearing this condition, as otherwise data would have been less consistent across regions in matching the diagnosis-based figures from GP clinical databases. Heart failure

was underestimated by GP data, although non significantly in the majority of the regions. Underestimation was highest in the oldest age band available in both data sources (85-95), where the prevalence is highest. This is consistent with the hypothesis that GPs belonging to HSDmight occasionally performless accurate data collection when visiting patients at home [27] or in residential care [22], or consider heart failure as a complication of other underlying conditions, such as ischaemic heart disease, rather than as a disease of its own. This would imply that the population detected by administrative data had a more severe form of the disease and was more often affected by disability. Another possibility is that administrative database overestimate prevalence because of lack of specificity of the case ascertainment algorithm. Indeed, according to a recent review of validated algorithms for case ascertainment of heart failure [28], algorithms using secondary discharge diagnosis showed lower positive predictive value (PPV) in several countries.

For COPD administrative data failed to detect the differences between regions that the other two sources consistently measured. Ascertainment of COPD from administrative sources has been shown to be challenging in other studies [29,30]. In this case, the algorithm detected a particular pattern of drug prescriptions, combining duration, intensity and ATC class, that had been identified through a consensus process in a group of experts that was reported in Anecchino et al. [21]. Although the pattern was specifically meant to avoid misclassification (e.g., with respect to asthma), it is possible that the conclusions of the study were in fact specific for the geographic area where the experts worked. In light of the limitations of the sampling design of our study, the overall good agreement with other data sources supports *a fortiori* validity of chronic disease surveillance using administrative data in the regions that were involved in the study. However, support for external validity of our results needs to be discussed. Althought we only collected data from few geographical areas, the same administrative data are available for the whole national population. We are in fact not claiming that administrative data from few geographically sparse areas can be used to estimate national prevalence of chronic diseases, but rather that administrative data seem to be consistently able to detect prevalence of some chronic diseases *around the area they were extracted from*. Our positive findings (treated diabetes, ischaemic heart disease, heart failure) are indeed probably due to the fact that typical health consumption patterns of such chronic patients are similar across regions. On the assumption that regions of the same macroarea of the country

(North, Center, South) are similar the one to the other to this respect, our data support the claim that estimates relying on the same algorithms should prove to be similarly effective. However, in some specific critical areas of the country where incomplete administrative data collection is suspected, a local evaluation is recommended.

Cohorts can be selected from administrative databases to perform population-based studies on patients with chronic diseases through further record-linkage with the same databases. This study cannot provide analytical tools to assess the limitations of the findings of such studies. However, no evidence emerges for major bias, except in the case of COPD, where regional differences with the other data sources are likely to be due to differences in the characteristics of the corresponding local cohorts.

Limitations

The first limitation of studies that make secondary use of existing healthcare data sources is that only prevalence of diagnosed cases is taken into account, and underestimation of actual population prevalence cannot be estimated [31]. An implicit assumption of both crude and adjusted rate estimation from administrative databases performed in this study was that PPV of the case detection was 100%, an assumption that we could not verify and that is not to be taken for granted, when, for instance, secondary discharge diagnosis or drug utilisation with no indication is used as a source of case ascertainment. Ecological validation studies cannot directly resolve this issue, as consistent ecological estimates between a data source and a reference gold standard might as well be due to coincidental inclusion of false positive and exclusion of false negative cases. Only validation studies perfomed using individual-level comparison with a gold standard could assess PPV and sensitivity.

CONCLUSION

This study supports the use of data from Italian administrative databases to estimate geographic differences in population prevalence of ischaemic heart disease, treated diabetes, diabetes mellitus and heart failure. The algorithm for COPD used in this study requires further refinement.

REFERENCES

1. Jutte DP, Roos LL, Brownell MD: Administrative record linkage as a tool for public health research. Annu Rev Public Health 2011, 32:91–108. [http://www.ncbi.nlm.nih.gov/pubmed/21219160]. [PMID: 21219160].
2. Moore DF, Lix LM, Yogendran MS, Martens P, Tamayo A: Stroke surveillance in Manitoba, Canada: estimates from administrative databases. Chronic Diseases in Canada 2008, 29:22–30. [http:// www.ncbi. nlm.nih.gov/pubmed/19036220]. [PMID: 19036220].
3. Wigertz A, Westerling R: Measures of prevalence: which healthcare registers are applicable? Scand J Public Health 2001, 29:55–62. [http:// www.ncbi.nlm.nih.gov/pubmed/11355718]. [PMID: 11355718].
4. Rector TS, Wickstrom SL, Shah M, Thomas Greeenlee N, Rheault P, Rogowski J, Freedman V, Adams J, Escarce JJ: Specificity and Sensitivity of Claims-Based, Algorithms for Identifying Members of Medicare+Choice Health Plans That Have ChronicMedical Conditions. Health Services Res 2004, 39(6 Pt 1):1839–1858. [PMID: 1553 3190 PMCID: 1361101].
5. Renard LM, Bocquet V, Vidal-Trecan G, Lair M-L Couffignal, S, Blum-Boisgard C: An algorithm to identify patients with treated type 2 diabetes usingmedico-administrative data. BMC Medical Informatics and Decision Making 2011, 11:23. [PMID: 21492480 PMCID: 3090314].
6. Hernán MA: With great data comes great responsibility: publishing comparative effectiveness research in epidemiology. Epidemiol (Cambridge,Mass.) 2011, 22(3):290–291. [http://www.ncbi.nlm. nih.gov/ pubmed/21464646]. [PMID: 21464646].
7. Cooke CR, Joo MJ, Anderson SM, Lee TA, Udris EM, Johnson E, Au DH: The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease. BMC Health Services Res 2011, 11:37. [http://www.ncbi.nlm.nih.gov/pubmed/ 21324188]. [PMID: 21324188].
8. Solberg LI: Are Claims Data Accurate Enough to Identify Patients for PerformanceMeasures or Quality Improvement? The Case of Diabetes, Heart Disease, and Depression. Am JMed Qual 2006, 21:238–245. [http://ajm.sagepub.com/cgi/doi/10.1177/ 1062860606288243].
9. Ellekjaer H, Holmen J, Kr ¨uger O, Terent A: Identification of incident stroke in Norway: hospital discharge data compared with a population-based stroke register. Stroke; a J Cerebral Circulation 1999, 30:56–60. [http://www.ncbi.nlm.nih.gov/pubmed/9880388]. [PMID: 9880388].
10. Lix LM, Yogendran MS, Shaw SY, Burchill C, Metge C, Bond R: Population-based data sources for chronic disease surveillance. Chronic Diseases in Canada 2008, 29:31–38. [http://www.ncbi.nlm. nih. gov/pubmed/19036221]. [PMID: 19036221].
11. Maio V, Yuen E, Rabinowitz C, Louis D, Jimbo M, Donatini A, Mall S, Taroni F: Using pharmacy data to identify those with chronic conditions in Emilia Romagna, Italy. J Health Services Res & Policy 2005, 10(4):232–238. [http://www.ncbi.nlm.nih.gov/pubmed/16259690]. [PMID: 16259690].
12. Lo Scalzo A, Donatini A, Orzella L, Cicchetti A, Profili S, Maresso A: Italy: Health system review. Health Syst Transition 2009, 11(6):1–216. 13. Chini F, Pezzotti P, Orzella L, Borgia P, Guasticchi G: Can we use the pharmacy data to estimate the prevalence of chronic conditions? a comparison of multiple data sources. BMC Public Health 2011, 11:688. [PMID: 21892946 PMCID: 3223740].
14. Capture-recapture andmultiple-record systems estimation I: History and theoretical development. International Working Group for Disease Monitoring and Forecasting. Am J Epidemiol 1995, 142(10):1047–1058. [http://www.ncbi.nlm.nih.gov/pubmed/7485050]. [PMID: 7485050].
15. Capture-recapture andmultiple-record systems estimation II: Applications in human diseases. International Working Group for Disease Monitoring and Forecasting. Am j epidemiol 1995, 142(10):1059–1068. [http://www.ncbi.nlm.nih.gov/pubmed/7485051]. [PMID: 7485051].
16. Giarrizzo ML, Pezzotti P, Silvestri I, Di Lallo D: [Estimating prevalence of diabetes mellitus in a Lazio province, Italy, by capture-recapture models]. Epidemiologia e prevenzione 2007, 31(6):333–339. [http://www. ncbi.nlm.nih.gov/pubmed/18326425]. [PMID: 18326425].
17. Gnavi R, Karaghiosoff L, Costa G, Merletti F, Bruno G: Socio-economic differences in the prevalence of diabetes in Italy: The populationbased Turin study. Nutr,Metab Cardiovasc Diseases 2008, 18(10):678–682. [http://www.sciencedirect.com/science/article/pii/S0939475307002049].

18. Costa G, Ricciardi G, Paci E(Eds): [United Italy, 150 years later: has equity in health and healthcare improved?]. In Epidemiologia e Prevenzione; 2011,35(5-6 Suppl 2) 1–134.
19. ATC: Structure and principles. [http://www.whocc.no/atc/ structure and principles/]. [Accessed: September 2012].
20. Simonato L, Baldi I, Balzi D, Barchielli A, Battistella G, Canova C, Cesaroni G, Corrao G, Collini F, Conti S, Costa G, Demaria M, Fornari C, Faustini A, Galassi C, Gnavi R, Inio A, Madotto F, Migliore E, Minelli G, Pellizzari M, Protti M, Romanelli A, Russo A, Saugo M, Tancioni V, Tessari R, Vianello A, Vigotti M(Eds): [Objectives, tools and methods for an epidemiological use of electronic health archives in various areas of Italy]. Epidemiologia e Prevenzione 2008,32(3 Suppl), 1–134. [http:// www.ncbi. nlm.nih.gov/pubmed/18928238].
21. Anecchino C, Rossi E, Fanizza C, De Rosa M, Tognoni G, Romero M: Prevalence of chronic obstructive pulmonary disease and pattern of comorbidities in a general population. Int J Chronic Obstructive Pulmonary Disease 2007, 2(4):567–574. [PMID: 18268930 PMCID: 2699,968].
22. Filippi A, Vanuzzo D, Bignamini AA, Sessa E, Brignoli O, Mazzaglia G: Computerized general practice databases provide quick and cost-effective information on the prevalence of angina pectoris. Italian Heart J: Official J Italian Federation of Cardiology 2005, 6:49–51. [http://www.ncbi. nlm.nih.gov/pubmed/15773273]. [PMID: 15773273].
23. [Demography in figures]. [http://demo.istat.it]. [Accessed: October 2011].
24. Cricelli C, Mazzaglia G, Samani F, Marchi M, Sabatini A, Nardi R, Ventriglia G, Caputi AP: Prevalence estimates for chronic diseases in Italy: exploring the differences between self-report and primary care databases. J Public HealthMed 2003, 25(3):254–257. [http://www.ncbi. nlm.nih. gov/pubmed/14575204]. [PMID: 14575204].
25. Leikauf J, Federman AD: Comparisons of self-reported and chart-identified chronic diseases in inner-city seniors. J AmGeriatrics Soc 2009, 57(7):1219–1225. [http://www.ncbi.nlm.nih.gov/ pubmed/ 19486197]. [PMID: 19486197].
26. Harris SB, Glazier RH, Tompkins JW, Wilton AS, Chevendra V, Stewart MA, Thind A: Investigating concordance in diabetes diagnosis between primary care charts (electronic medical records) and health administrative data: a retrospective cohort study. BMC Health Services Res 2010, 10:347. [PMID: 2118 2790 PMCID: 3022877].
27. Filippi A, Sessa E, Pecchioli S, Trifir G, Samani F, Mazzaglia G: Homecare for patients with heart failure in Italy. Italian Heart J: Official J Italian Federation of Cardiology 2005, 6(7):573–577. [http:// www.ncbi.nlm.nih. gov/pubmed/16274019]. [PMID: 16274019].
28. Saczynski JS, Andrade SE, Harrold LR, Tjia J, Cutrona SL, Dodd KS, Goldberg RJ, Gurwitz JH: A systematic review of validated methods for identifying heart failure using administrative data. Pharmacoepidemiology and Drug Saf 2012, 21:129–140. [http:// onlinelibrary.wiley.com/ doi/10.1002/pds.2313/abstract].
29. Lacasse Y, Montori VM, Lanthier C, Maltis F: The validity of diagnosing chronic obstructive pulmonary disease from a large administrative database. Can Respir J: J Canadian Thoracic Soc 2005, 12(5):251–256. [http://www.ncbi.nlm.nih.gov/pubmed/16107913]. [PMID: 16107913].
30. Faustini A, Canova C, Cascini S, Baldo V, Bonora K, De Girolamo G, Romor P, Zanier L, Simonato L: The reliability of hospital and pharmaceutical data to assess prevalent cases of chronic obstructive pulmonary disease. COPD 2012, 9(2):184–196. [PMID:, 2240 9483].
31. Prevalence of chronic diseases in older Italians: comparing self-reported and clinical diagnoses. The Italian Longitudinal Study on AgingWorking Group. Int J Epidemiol 1997, 26(5):995–1002. [http:// www.ncbi.nlm.nih.gov/pubmed/9363520]. [PMID: 9363520].

CHAPTER 2

Automatic identification of type 2 diabetes, hypertension, ischaemic heart disease, heart failure and their levels of severity from Italian General Practioners' electronic medical records: a validation study

Rosa Gini [1,2], Martijn J Schuemie [3,4], Giampiero Mazzaglia [5], Francesco Lapi [5], Paolo Francesconi [1], Alessandro Pasqua [5], Elisa Bianchini [5], Carmelo Montalbano [6], Giuseppe Roberto [1], Valentina Barletta [1], Iacopo Cricelli [6], Claudio Cricelli [5], Giulia Dal Co [7], Mariadonata Bellentani [7], Miriam Sturkenboom [2], Niek Klazinga [8]

1. Agenzia regionale di sanità della Toscana, Osservatorio di epidemiologia; 50141 Florence, Italy
2. Department of Medical Informatics, Erasmus Medical Center; 3015 GJ Rotterdam, The Netherlands
3. Janssen Research & Development, Epidemiology; Titusville, New Jersey, United States
4. Observational Health Data Sciences and Informatics (OHDSI); New York, New York, United States
5. Health Search, Italian College of General Practitioners and Primary Care; 50100 Florence, Italy
6. Genomedics; 50100 Florence, Italy
7. Agenzia Nazionale per i Servizi Sanitari Regionali; 00100 Rome, Italy
8. Academic Medical Center, University of Amsterdam; 1100 DD Amsterdam, The Netherlands

Submitted for publication

ABSTRACT

Objectives
The Italian project MATRICE aimed to assess how well cases of type 2 diabetes (T2DM), hypertension, ischaemic heart disease (IHD) and heart failure (HF) and their levels of severity can be automatically extracted from the Health Search/CSD Longitudinal Patient Database (HSD). From the medical records of the general practitioners (GP) who volunteered to participate, cases were extracted by algorithms based on diagnosis codes, keywords, drug prescriptions and results of diagnostic tests. A random sample of identified cases was validated by interviewing their GPs.

Setting
HSD is a database of primary care medical records. A panel of 12 GPs participated in this validation study.

Participants
300 patients were sampled for each disease, except for HF, where 243 patients were assessed.

Outcome measures
Positive predictive value (PPV) was assessed for the presence/absence of each condition against the GP's response to the questionnaire, and Cohen's kappa was calculated for agreement on the severity level.

Results
The PPV was 100% [99-100] for T2DM and hypertension, 98% [96-100] for IHD and 55% [49-61] for HF. Cohen's kappa for agreement on the severity level was 0.70 for T2DM and 0.69 for both hypertension and IHD.

Conclusions
This study shows that subjects with T2DM, hypertension or IHD can be validly identified in HSD by automated identification algorithms. Automatic queries for levels of severity of the same diseases compare well with the corresponding clinical definitions, but some misclassification occurs. For HF further research is needed to refine the current algorithm.

Italy is facing an increasing burden of chronic health conditions due to aging of the population. To provide adequate and fair health care across regions Italy was advised by the Organisation for Economic Co-operation and Development to develop a set of standards around the processes and outcomes of primary care, and to develop a national quality governance model to support regions in delivering care of uniform quality across the country [OECD2015]. The Italian National Agency for Regional Healthcare Services started the MATRICE Project in 2011. MATRICE was aimed at developing tools to compare quality of healthcare across Italian regions of four chronic diseases: type 2 diabetes (T2DM), hypertension, ischaemic heart disease (IHD) and heart failure (HF). One of the objectives was to assess the validity of routine care data to monitor quality of healthcare supply [AGENAS2016, ARS2016].

The Health Search IMS Health Longitudinal Patient Database (HSD) is a longitudinal primary care medical record database that was set up by members of the Italian College of General Practitioners (SIMG). More than 900 physicians, uniformly distributed across Italy, share their de-identified clinical records in the HSD. These data are extensively used for epidemiological and public health research [Cricelli2003, Savica2007, Sacchetti2008, Mazzaglia2009, Ravera2009, Cazzola2010, Cazzola2013, Sultana2014, HSD2016].

The HSD database is very similar to other Primary Care databases: for instance, in the United Kingdom, the Clinical Practice Research Datalink (CPRD, formerly GPRD) [Walley1997], The Health Information Network (THIN) [Lewis2007] and QResearch [Hippisley-Cox2004]; in Canada, the Canadian Primary Care Sentinel Surveillance Network (CPCSSN); in the Netherlands, the Integrated Primary Care Information database (IPCI) [Vlug1999]; in Spain, the database Base de datos para la Investigación Farmacoepidemiológica en Atención Primaria (BIFAP) [SalvadorRosa2002]. A common feature of these countries is that the GP serves a well-defined population and is the gatekeeper to secondary care.

In this type of medical records every visit is recorded and all diagnoses, prescriptions and measurements are recorded as part of a general practitioner's daily practice. Moreover, information from specialist referrals is reported back to the general practitioner (GP) and stored in the same medical records. The medical records replace the paper records that once existed and may be considered a rather comprehensive list of health problems requiring care. Because of its longitudinal, population-based nature, this type of databases serves well for many research purposes, such as estimate of burden of disease,

parmacoepidemiologic and health services research. Since information is collected primarily for the provision of care, the quality of coding diagnoses may not always be accurate and also this varies by type of disease [Jordan2004, Kadhim-Saleh2013]. Moreover, the same code may be used for different clinical definitions of a disease, when diagnostic standards are not uniform across healthcare communities or change over time. As a consequence, case-finding algorithms that retrieve subjects from such data sources using diagnostic codes may unintentionally retrieve subjects whose clinical condition does not correspond to the one intended for a study. On the other hand, GPs may record clinical conditions as a free text note, and a retrieval strategy using only coded diagnoses may miss some cases. As a result, algorithms using other sources, such as results from laboratory tests, have been developed to query this type of data sources [Tu2011]. Disease-specific validation studies of case-finding algorithms in GP medical record databases have been performed in CPRD [Herrett2010, Khan2010], CPCSSN [Kadhim-Saleh2013, Williamson2014], IPCI and HSD itself [Coloma2013, Valkhoff2014].

A panel of experts in the MATRICE Project established precise clinical definitions of T2DM, hypertension, IHD and HF, and identified levels of severity of the conditions that had to be distinguished for the purpose of monitoring healthcare quality. The aim of this study was to estimate the positive predictive value of case-finding algorithms to detect such conditions and levels of severity from the GP medical records collected by HSD, against a gold standard based on manual comparison with the clinical definitions chosen by the MATRICE panel.

METHODS

Setting

Italy has a tax-based, universal coverage national health system. Every Italian resident is entitled to choose a GP, although parents might instead opt for a specialist paediatrician for their children, up to the age of 15. Therefore, each resident from the age of 16 onward is specifically registered with a GP. GPs are the "gatekeepers" of the system, meaning that patients can only access secondary care within the healthcare system upon referral of their GP [LoScalzo2009, OECD2015]. Secondary care is accessed either free of charge or upon a small copayment.

During their daily practice, GPs record all clinical findings, diagnoses and prescriptions in their electronic medical records. GPs participating in HSD all use the same software, which requires that each prescription is associated with a specific disease code. A disease code may be labelled as 'suspect' when further clinical ascertainment is needed. Results from laboratory tests may be recorded as well. Moreover, free-text fields are available in the software to collect clinical notes on diagnoses, signs, symptoms and referral letters from specialists or from hospitals. Every 6 months GPs send their data to a central repository, after anonymization. The central repository performs quality controls, like estimation of prevalence of common diseases, and selects GPs whose data prove to be accurate [Cricelli2003]. Currently, data of 700 out of 900 GPs, uniformly distributed across Italy, are considered accurate according to data quality checking [HSD2016].

Clinical definition of the diseases and of their levels of severity

A panel of cardiologists, diabetologists, epidemiologists and experts in organization of primary care services participating in the MATRICE Project first established clinical definitions of the four diseases and of their levels of severity. The levels of severity were selected according to whether national and international clinical guidelines contained specific indications for treatment or diagnostic follow-up in the patients with that condition. For instance, a patient with IHD after an episode of acute myocardial infarction (AMI) has an indication for treatment with beta-blockers [Fihn2012], hence history of AMI is a relevant level of severity for IHD.

The detailed definitions of the diseases and of the levels of severity are depicted in Table 1. For T2DM, the clinical definition was at least two abnormal measurements among fasting plasma glucose, or two-hour plasma glucose after a load of glucose, or glycated haemoglobin; or just one abnormal measurement of plasma glucose if symptoms of hyperglycaemia were observed. Four levels of severity of the disease were identified, according to presence/absence of indication for insulin and presence/absence of complications or organ damage. For hypertension, diagnostic criteria were two abnormal measurements for both systolic and diastolic blood pressure confirmed either by a Holter blood pressure measurement or by home blood pressure monitoring. Three levels of severity were identified: no organ damage or diabetes or stroke; organ damage or diabetes or stroke without HF; hypertension with HF. For IHD the

Table 1. **Clinical definition of diseases and levels of severity. ACC/AHA: American Cardiology Association and American Heart Association.**

Type 2 diabetes mellitus	
Clinical definition	**Levels of severity**
Syndrome diagnosed on the basis of the following criteria outlined in a first test and confirmed with a second test in a adult, non-pregnant patient, without typical symptoms of the disease: Fasting plasma glucose ≥ 126 mg / dl (no caloric intake for at least 8 hours), or two-hour plasma glucose ≥ 200 mg / dl during an OGTT after a load of 75 g glucose, or glycated hemoglobin ≥ 6.5%. Or, in a patient with classic symptoms of hyperglycemia or hyperglycemic crisis, a random plasma glucose of plasma glucose ≥ 200 mg/dl (regardless of food intake).	LEVEL 1 Clinical definition of the disease, no indication for insulin therapy and no of the complications listed in level 3 LEVEL 2 Clinical definition of the disease, indication for insulin therapy and absence of complications listed in level 3 LEVEL 3 Clinical definition of the disease, no indication for insulin therapy and one of the following:: (1) Arterial stenosis (coronary, carotid, peripheral arteries of lower extremities), angina pectoris, MI, TIA, ischemic stroke of atherosclerotic origin, intermittent claudication, diabetic foot ulcer, lower limb amputation (2) Retinopathy (3) incipient diabetic nephropathy (microalbuminuria) or overt (albuminuria or GFR abnormal) / Dialysis LEVEL 4 As in level 3, except that insulin is indicated
Hypertension	
Clinical definition	**Levels of severity**
Syndrome characterized by arterial systolic blood pressure above 140 mmHg and / or diastolic blood pressure above 90 mmHg in at least two measurements (patient at rest) confirmed by Holter blood pressure measurements or by home blood pressure monitoring (2 measurements in the morning and two in the evening for seven days and then calculating the average of all measurements after discarding those of the first day (as recommended by the ESH guidelines)	LEVEL 1 Clinical definition of the disease, absence of organ damage and of diabetes LEVEL 2 Clinical definition of the disease, no HF in level at least C of the ACC/AHA classification and at least one of the following conditions: type 2 diabetes; hypertrophy (ECG o Echo), dilatation or left ventricular asyergy (Echo); hypertensive retinopathy; GFR abnormal; microalbuminuria or proteinuria; atherosclerotic plaques in carotid arteries; atherosclerotic peripheral arterial occlusive disease; angina pectoris; coronary revascularization; AMI; TIA or ischemic stroke due to atherosclerosis; hypertensive encephalopathy; abdominal aortic aneurysm; aortic dissection; cerebral hemorrhage LEVEL 3 Clinical definition of the disease and HF in stage C or D of the ACC/AHA classification
Ischaemic heart disease	
Clinical definition	**Levels of severity**
Clinical syndrome characterized by typical angina chest pain, and / or transient myocardial ischemia verified by stress ECG or imaging, and / or significant coronary arteries occlusion verified with angiography, or history of previous AMI.	LEVEL 1 Clinical definition of the disease, no evidence of previous AMI nor PTCA, no evidence of HF in stage C or D of the ACC/AHA classification LEVEL 2 Evidence of previous PTCA, no evidence of previous AMI, no evidence of HF in stage C or D of the ACC/AHA classification LEVEL 3 Evidence of previous AMI, no evidence of previous PTCA, no evidence of HF in stage C or D of the ACC/AHA classification LEVEL 4 Evidence of previous PTCA and AMI, no evidence of HF in stage C or D of the ACC/AHA classification LEVEL 5 Clinical definition of the disease and evidence of HF in stage C or D of the ACC/AHA classification

Table 1. **Continued**

Heart failure	
Clinical definition	**Levels of severity**
Stage C or D of the ACC/AHA classification: syndrome characterized by the presence of symptoms and signs, current or prior dyspnea and / or fatigue and / or fluid retention (peripheral edema and / or pulmonary stasis), and the presence of structural heart disease (left ventricular systolic dysfunction with ejection fraction (EF) <50% and / or left ventricular diastolic dysfunction and / or right ventricular dysfunction) detected by echocardiography	*None*

clinical definition referred to symptoms (angina pain) or to history of acute myocardial infarction or to bioimaging observation of coronary ischemia. Five levels of severity of IHD were identified: the most severe was HF, and among those free from HF presence/absence of history of AMI and presence/absence of percutaneous transluminal coronary angioplasty (PTCA) classified the four levels. HF was also identified as a condition in its own, and the clinical definition was stage C or D of the classification by the American College of Cardiology and of the American Heart Association [Hunt2005].

Case identification in the primary care medical records

A panel comprising epidemiologists from HSD and GPs belonging to SIMG, with expertise in the clinical areas of interest, developed ad hoc algorithms to identify from the GP medical records cases matching the clinical definitions of the MATRICE Project. In each algorithm the inclusion and exclusion criteria of the clinical definition were mapped to a list of ICD9CM codes or, when deemed necessary, to free text and to conditions on diagnostic tests.

Case validation

An invitation to participate in the validation study was circulated by SIMG to a sample of GPs, and participation was voluntary.

A data collection plugin for the medical record software was developed by HSD and installed in the computers of the GPs who accepted to participate in the study.

For each disease the plugin applied the algorithm to the whole list of active patients in the date of data collection, and selected a random sample of 25

patients from the resulting list of cases. The plugin then showed the names of the patients in the sample to the GP for assessment. In the same screen, the clinical definitions of the disease and of its levels of severity were presented. The GP was aware that the cases were selected for a specific condition among T2DM, hypertension, IHD and HF, but was blinded to the level of severity. The GP had the choice of indicating that the patient was not affected by the disease (false positive: FP), or that the patient was affected but the level of severity could not be assessed on the basis of the information available to the GP (not staged: NS), or to assign a level. In the process, the GP was free to access the patient's medical record. Figure 1 shows a screenshot of the data collection plugin (patients listed are not real).
When the GP had completed the manual assessment of the 4 samples of patients, the plugin applied the algorithms for the level of severity, linked the new columns to the dataset resulting from the questionnaire, anonymized the final dataset and transmitted it to HSD for statistical analysis.
Data collection was performed in July 2013.

Statistical analysis
The formula to compute Positive Predictive Value (PPV), both for presence/absence of the condition and for each of the levels of severity, was

PPV= True Positives/(True Positives + False Positives)

and 95% confidence intervals (CI) were estimated. For the three diseases where levels of severity had been validated (T2DM, hypertension and IHD), Cohen's kappa was computed to the categorical distribution of levels of severity. Cohen's kappa discounts from observed concordance (the percentage of subjects who are classified in the same level by the algorithms and the GP) the expected concordance (the percentage of subjects who would be classified the same if assignment had been performed randomly) by means of the following formula

K = (Observed Concordance – Expected Concordance)/
(1 – Expected Concordance)

Cohen's kappa provides an overall measure of agreement about levels of severity. Analysis was performed using Stata 12.

RESULTS

Algorithms for primary care medical records

The algorithms detecting the diseases and levels of severity from the primary care medical records are listed in Table 2. Each algorithm consists of a sequence of rules, each acting as an inclusion criterion, a refinement criterion (linked to the inclusion criterion with the logical connector AND), or a refinement criterion (linked to the inclusion criterion with the logical connectors AND NOT). Each rule is itself composed by subqueries (represented in the table by keywords in round parenteses), and subjects matching at least one of the subqueries are included in the rule, that is, the subqueries are linked to each other with the logical connector OR. No specific temporal sequence between subqueries is requested. Every sub query selects records matching a specific list of codes, free text keywords and/or diagnostic test levels, which are listed in the Electronic Supplementary Material 1 (http://www.ars.toscana.it/files/progetti/informatica_medica/thematrix/ESM_1_revised.pdf). All the subqueries are applied to the whole set of longitudinal observations of the patients up to the index date, except when specified otherwise.

In all subqueries, records labelled with 'suspect' were excluded, except in the case of the subquery detecting patients with acute myocardial infarction, used to detect levels of severity 3 and 4 in patients with IHD.

Validation

A panel of 12 GPs participated in the validation study. A total of 300 patients were identified and validated for each disease, except for HF, where due to low prevalence of the condition only 243 patients were included.

The PPV of the algorithms were 100% (CI: 99-100) for T2DM and hypertension and 98% (CI: 96-100) for IHD. For HF PPV was 55% (CI: 49-61).

For T2DM the second and fourth levels of severity had very high PPV (88% and 93% respectively). Around the 20% of patients without indication for insulin (first and third level of severity) had their presence of complications misclassified. Both possibilities took place: patients with complications were identified as being free from them, and vice versa. Overall Cohen's kappa was 0.70, a good level of agreement (Table 3).

Among hypertensive patients every level of severity was misclassified in less than the 20% of patients, and in the case of the middle level (organ damage

and/or diabetes, no HF) patients were in fact almost all less severe with respect to the level they were automatically assigned to. Cohen's kappa was 0.69, showing good agreement (Table 3).

In the case of IHD the first and fifth levels of severity had excellent PPV, while AMI was incorrectly identified by the algorithm in 22% of cases. The two levels of severity characterized by presence of PTCA were never manually indicated by the GPs, and the automatic algorithm was in almost perfect agreement in both cases. Overall, Cohen's kappa was 0.69, showing good agreement (Table 3).

Table 2. Algorithms to detect diseases and levels of severity. The algorithms are described by means of subqueries, represented by keywords in upper case between parentheses, whose details are in Supplementary Table 1. In particular, measurements are as follows. (GFR): Glomerular filtration rate < mL/min/1.73 m^2 (BP): Systolic blood pressure>140 mmHg at least twice ever OR Diastolic blood pressure >90 mmHg at least twice ever (LVEF): Left ventricular ejection fraction< 50% (DM TESTS): plasma glucose >200mg/dl after an oral load of 75 g glucose OR fasting plasma glucose>=126 mg/dl OR glycated haemoglobin>=6.5%. In all subqueries records labelled with 'suspect' were excluded, except in the case of the subquery (AMI) (Acute myocardial infarction) used to detect levels of severity 3 and 4 of IHD.

Type 2 diabetes mellitus	
Algorithm for the disease	**Algorithms for levels of severity**
[(DM) OR (GFR) OR (DM TESTS)] **AND NOT** (DM1)	Algorithm for the disease **AND** (DM UNCOMPLICATED) **AND NOT** [(INSULIN) OR (DM2 CHRONIC COMPLICATIONS) OR (DM2 ASYMPTOMATIC COMPLICATIONS) OR (CVD)]
	Algorithm for the disease **AND** (DM UNCOMPLICATED) **AND** (INSULIN) **AND NOT** [(DM2 CHRONIC COMPLICATIONS) OR (DM2 ASYMPTOMATIC COMPLICATIONS) OR (CVD)]
	Algorithm for the disease **AND** [(DM CHRONIC COMPLICATIONS) OR (DM ASYMPTOMATIC COMPLICATIONS) OR (CVD)] **AND NOT** (INSULIN)
	Algorithm for the disease **AND** [(DM CHRONIC COMPLICATIONS) OR (DM ASYMPTOMATIC COMPLICATIONS) OR (CVD)] **AND** (INSULIN)
Hypertension	
Algorithm for the disease	**Algorithms for levels of severity**
(HYPERTENSION) OR (BP)	Algorithm for the disease **AND NOT** [LEVEL 2 OR LEVEL 3]
	Algorithm for the disease **AND** [(DM2) OR (HYPERTENSION COMPLICATIONS) OR (HYPERTENSIVE RETINOPATHY) OR (CHRONIC KIDNEY DISEASE) OR (ATHEROSCLEROTIC ARTERIOPATHY) OR (CAROTID ATHEROSCLEROSIS) OR (HYPERTENSIVE ENCEPHALOPATHY) OR (AORTIC ANEURYSM)] **AND NOT** [(HF) OR (LVEF)]

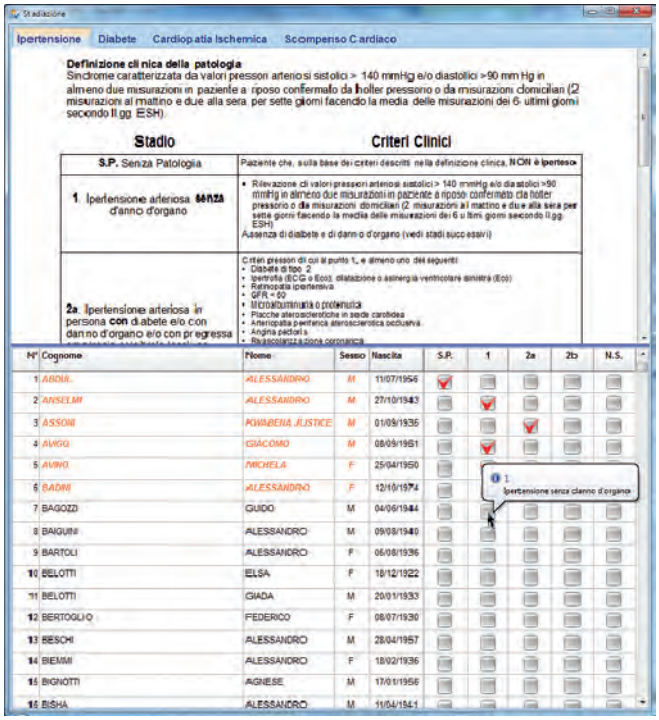
Table 2. Continued

	Algorithm for the disease **AND** [(HF) OR (LVEF) OR (VENTRICULAR DYSFUNCTION)]
Ischaemic heart disease	
Algorithm for the disease	**Algorithms for levels of severity**
(CHD)	Algorithm for the disease **AND** (CHD NO AMI) **AND NOT** [(AMI) OR (PAST AMI) OR (CORONARY REVASCULARIZATION ICD9CM) OR (CORONARY REVASCULARIZATION FREE TEXT) OR (HF) OR (LVEF) OR (VENTRICULAR DYSFUNCTION)]
	Algorithm for the disease **AND** [(CORONARY REVASCULARIZATION ICD9CM) OR (CORONARY REVASCULARIZATION FREE TEXT)] **AND NOT** [(AMI) OR (PAST AMI) OR (HF) OR (LVEF) OR (VENTRICULAR DYSFUNCTION)]
	Algorithm for the disease **AND** [(AMI) OR (PAST AMI)] **AND NOT** [(CORONARY REVASCULARIZATION ICD9CM) OR (CORONARY REVASCULARIZATION FREE TEXT) OR (HF) OR (LVEF) OR (VENTRICULAR DYSFUNCTION)]
	Algorithm for the disease **AND** [(AMI) OR (PAST AMI) OR (CORONARY REVASCULARIZATION ICD9CM) OR (CORONARY REVASCULARIZATION FREE TEXT)] **AND NOT** [(HF) OR (LVEF) OR (VENTRICULAR DYSFUNCTION)]
	Algorithm for the disease **AND** [(HF) OR (LVEF)]
Heart failure	
Algorithm for the disease	**Algorithms for levels of severity**
(HF) OR (LVEF) OR (VENTRICULAR DYSFUNCTION)	*None*

Table 3. Results of the validation.

Type 2 diabetes mellitus (N=300, true positives overall: 300)						
Level	Description	Automatically assigned	True positives	True level of false positives	PPV [95% CI]	Cohen's K
1	Diabetes with no evidence of organ damage nor complications and no indication for insulin therapy	129	101	Level 3: 28	0.78 [0.70-0.85]	0.70
2	Diabetes with indication for insulin but no evidence of organ damage nor complications	17	15	Level 4: 2	0.88 [0.64-0.99]	
3	Diabetes with evidence of organ damage or complications and no indication for insulin	127	104	Level 1: 23	0.82 [0.74-0.88]	
4	Diabetes with evidence of organ damage or complications and with indication for insulin	27	25	Level 3: 2	0.93 [0.76-0.99]	
Hypertension (N=300, true positives overall: 300)						
Level	Description	Automatically assigned	True positives	True level of false positives	PPV [95% CI]	Cohen's K
1	Hypertension with no organ damage nor diabetes	138	119	Level 2: 18 Level 3: 1	0.86 [0.79-0.91]	0.69
2	Hypertension with organ damage and/or stroke and/or diabetes, no HF	155	127	Level 1: 28	0.82 [0.75-0.88]	
3	Hypertension and HF	7	6	Level 2: 1	0.86 [0.42-0.99]	
Ischaemic heart disease (N=300, true positives overall: 298)						
Level	Description	Automatically assigned	True positives	True level of false positives	PPV [95% CI]	Cohen's K
1	IHD, no AMI, no HF, no PTCA	56	54	No IHD: 2	0.96 [0.88-0.99]	0.71
2	IHD with PTCA, no AMI, no HF	1	0	Level 5: 1	0	
3	IHD with AMI, no PTCA, no HF	195	152	Level 1: 43	0.78 [0.71-0.84]	
4	IHD with PTCA and AMI, no HF	0	0	-	-	
5	IHD with HF	48	45	Level 3: 3	0.94 [0.83-0.99]	
Heart failure (N=243, true positives: 134)						

Figure 1. Screenshot of the data collection plugin. The GP participating to the study was asked to fill in the four tables, one per disease. In the table "S.P." meant "without the disease" and "N.S." meant that the patient had the disease but the GP was not able to assess the level of severity. The clinical definition of the condition and of each level of severity could be browsed in the upper part of the screen, while, whenever the pointer stopped on a cell in the table, a tooltip suggested the label of the corresponding level. The patients listed in the figure are not real.



DISCUSSION

This study shows that almost all of the automatically detected cases of T2DM, hypertension and IHD, but only the 55% of cases of HF were true cases as assessed by the GP, on the basis of their own records and personal knowledge. Automatic classification of levels of severity of T2DM, hypertension and IHD was acceptable although less accurate. In the case of IHD mild cases were misclassified as severe, while in T2DM and hypertension both possibilities took place: severe patients were automatically classified as mild and vice versa. Our results provide guidance to the interpretation of results of studies using those algorithms to define variables in medical records or HSD, for instance to monitor quality of healthcare.

The excellent PPV of some algorithms is not unexpected: algorithms to detect diabetes (irrespective of type) and hypertension, as well as other chronic conditions, had similarly high validity in the CPCSSN [Kadhim-Saleh2013, Williamson2014]. The low PPV that we observed in the case of HF is unsurprising as well. Indeed, HF is a syndrome, and several different clinical definitions of HF have been used among clinicians in the recent past, such as the Framingham and European Society of Cardiology criteria. It has been shown that changing definition has a noticeable impact on the epidemiology of the condition [DiBari2004]. Our definition comprises two of the four stages of the classification of the American College of Cardiology and of the American Heart Association, which has been itself revised repeatedly in recent years [Hunt2005]. It is likely that the GPs in the sample themselves adopt a different definition to diagnose HF with respect to the one proposed by the MATRICE panel. Further research could investigate whether a clinical definition modelled on the Italian guidelines may be more easily identified in primary care medical records. Finally, poor performance of ICD9CM codes in identifying a specific clinical definition of HF has been consistently reported in literature [Quach2010].

The results we observed for levels of severity are probably due to reasons which are more specific to HSD. In the case of T2DM, some misclassification occurred between Level 1 and Level 3, that is, patients with or without complications but without indication for insulin use. Adding rules that explore free text comments to the diagnostic codes may improve those algorithms. In the algorithm developed for this study, patients with IHD labelled with 'suspect' AMI were included in the 'AMI' level of severity, because in a previous study

specific algorithms to identify AMI had been observed to have low sensitivity in HSD [Coloma2013]. As a result, our sensitive algorithm did indeed capture all the cases of AMI in patients with IHD; however, 22% of patients had not had an AMI, so this strategy needs to be reconsidered for future research. Remarkably, GPs were not aware of a single case of PTCA in their patients, as confirmed by automatic querying their records. The absence of PTCA could be due to the fact that this procedure is only performed in hospital. In Italy, hospitals do not provide GPs with discharge letters, rather patients themselves must describe to the GP what happened during an inpatient care episode, and the PTCA procedure may be communicated inadequately to the GP.

Strategies to improve communication between hospital and primary care should be implemented in Italy, not only for the purpose of improving quality of primary care medical records, but also to improve healthcare for patients with severe cardiovascular conditions.

Limitations

The sample of GPs was self-selected and may have been composed of those with more accurate data recording attitudes. In particular, 2 GPs in the validation sample also participated in the panel that created the algorithms. Positive predictive value of the algorithms may be lower in the general group of GPs contributing to HSD.

Providing a direct estimate of sensitivity of the case-finding algorithms was not an aim of this study, because it would have been unfeasible for GPs to assess a large enough sample of their patients. Indeed, the number of patients needed to estimate sensitivity of a test is bigger with less prevalent conditions: for instance, to estimate sensitivity of a case-finding algorithm for T2DM with a marginal error of 5%, even assuming 10% prevalence (an overestimation, according to common estimates in Italy [Gini2013]) and 95% sensitivity, would have requested an additional sample of more than 700 subjects; to obtain a valid estimate of sensitivity for HF, which has much lower prevalence and a lower expected sensitivity, GPs would have needed to assess thousands of subjects [Hajian-Tilaki2014].

CONCLUSIONS

This study shows that subjects with T2DM, hypertension or IHD can be validly identified in HSD by automated identification algorithms. Automatic queries for levels of severity of the same diseases compare well with the corresponding clinical definitions, but some misclassification occurs. For HF further research is needed to refine the current algorithm.

REFERENCES

[AGENAS2016] Agenzia nazionale per i Servizi Sanitari Regionali. Programma Mattoni del SSN - Progetto MATRICE. http://www.agenas.it/ images/agenas/In%20primo%20piano/Matrice/Progetto_MATRICE_Scheda_informativa.pdf. Accessed August 2016 [Italian]

[ARS2016] Agenzia regionale di sanità della Toscana. Data integration for chronic diseases management in outpatient settings (MATRICE Project). https://www.ars.toscana.it/en/project/ chronic-diseases/2460-matrice-project.html. Accessed August 2016

[Cazzola2010] Cazzola M, Bettoncelli G, Sessa E, Cricelli C, Biscione G. Prevalence of comorbidities in patients with chronic obstructive pulmonary disease. Respiration. 2010;80(2):112–9.

[Cazzola2013] Cazzola M, Calzetta L, Lauro D, Bettoncelli G, Cricelli C, Di Daniele N, et al. Asthma and COPD in an Italian adult population: role of BMI considering the smoking habit. Respir Med. 2013;107(9):1417–22.

[Coloma2013] Coloma PM, Valkhoff VE, Mazzaglia G, Nielsson MS, Pedersen L, Molokhia M, et al. Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems: a validation study in three European countries. BMJ Open. 2013

[Cricelli2003] Cricelli C, Mazzaglia G, Samani F, Marchi M, Sabatini A, Nardi R, et al. Prevalence estimates for chronic diseases in Italy: exploring the differences between self-report and primary care databases. J Public Health Med. 2003;25(3):254–7.

[DiBari2004] Di Bari M, Pozzi C, Cavallini MC, Innocenti F, Baldereschi G, De Alfieri W, et al. The diagnosis of heart failure in the community. Comparative validation of four sets of criteria in unselected older adults: the ICARe Dicomano Study. J Am Coll Cardiol. 2004;44(8):1601–8.

[Fihn2012] Fihn SD, Gardin JM, Abrams J, Berra K, Blankenship JC, Dallas AP, et al. 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS Guideline for the Diagnosis and Management of Patients With Stable Ischemic Heart Disease: Executive Summary. A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, and the American College of Physicians, American Association for Thoracic Surgery, Preventive Cardiovascular Nurses Association, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons. Circulation. 2012;126(25):3097–137.

[Gini2013] Gini R, Francesconi P, Mazzaglia G, Cricelli I, Pasqua A, Gallina P, et al. Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey. BMC Public Health. 2013;13(1):15.

[Hajian-Tilaki2014] Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. Journal of Biomedical Informatics. 2014;48:193–204.

[Herrett2010] Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. British Journal of Clinical Pharmacology. 2010;69(1):4–14.

[Hippisley-Cox2004] Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: a new general practice database for research. Inform Prim Care. 2004;12(1):49–50.

[HSD2016] Health Search Web page. https://www.healthsearch.it/?lang=en Accessed August 2016

[Hunt2005] Hunt SA, Abraham WT, Chin MH, Feldman AM, Francis GS, Ganiats TG, et al. ACC/AHA 2005 Guideline Update for the Diagnosis and Management of Chronic Heart Failure in the Adult A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Update the 2001 Guidelines for the Evaluation and Management of Heart Failure): Developed in Collaboration With the American College of Chest Physicians and the International Society for Heart and Lung Transplantation: Endorsed by the Heart Rhythm Society. Circulation. 2005;112(12):e154–235.

[Kadhim-Saleh2013] Kadhim-Saleh A, Green M, Williamson T, Hunter D, Birtwhistle R. Validation of the Diagnostic Algorithms for 5 Chronic Conditions in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): A Kingston Practice-based Research Network (PBRN) Report. J Am Board Fam Med. 2013;26(2):159–67.

[Khan2010] Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. Br J Gen Pract. 2010;60(572):e128–e136.

[Jordan2004] Jordan K, Porcheret M, Croft P. Quality of morbidity coding in general practice computerized medical records: a systematic review. Fam Pract. 2004;21(4):396–412.

[Lewis2007] Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. Pharmacoepidemiol Drug Saf. 2007;16(4):393–401.

[Mazzaglia2009] Mazzaglia G, Ambrosioni E, Alacqua M, Filippi A, Sessa E, Immordino V, et al. Adherence to Antihypertensive Medications and Cardiovascular Morbidity Among Newly Diagnosed Hypertensive Patients. Circulation. 2009;120(16):1598–605.

[OECD2015] OECD. OECD Reviews of Health Care Quality. Italy 2014: Raising Standards. OECD Publishing; 2015.

[Quach2010] Quach S, Blais C, Quan H. Administrative data have high variation in validity for recording heart failure. Can J Cardiol. 2010;26(8):e306–e312.

[Ravera2009] Ravera M, Noberasco G, Re M, Filippi A, Gallina AM, Weiss U, et al. Chronic kidney disease and cardiovascular risk in hypertensive type 2 diabetics: a primary care perspective. Nephrol Dial Transplant. 2009;24(5):1528–33.

[Sacchetti2008] Sacchetti E, Trifirò G, Caputi A, Turrina C, Spina E, Cricelli C, et al. Risk of stroke with typical and atypical anti-psychotics: a retrospective cohort study including unexposed subjects. J Psychopharmacol (Oxford). 2008;22(1):39–46.

[SalvadorRosa2002] Salvador Rosa A, Moreno Pérez JC, Sonego D, García Rodríguez LA, Iglesias A, De FJ. El Proyecto BIFAP: Base de datos para la Investigación Farmacoepidemiológica en. Aten Primaria. 2002;30(10):655–61.

[Savica2007] Savica R, Beghi E, Mazzaglia G, Innocenti F, Brignoli O, Cricelli C, et al. Prescribing patterns of antiepileptic drugs in Italy: a nationwide population-based study in the years 2000-2005. Eur J Neurol. 2007;14(12):1317–21.

[Sultana2014] Sultana J, Italiano D, Spina E, Cricelli C, Lapi F, Pecchioli S, et al. Changes in the prescribing pattern of antidepressant drugs in elderly patients: an Italian, nationwide, population-based study. Eur J Clin Pharmacol. 2014;70(4):469–78.

[Tu2011] Tu K, Manuel D, Lam K, Kavanagh D, Mitiku TF, Guo H. Diabetics can be identified in an electronic medical record using laboratory tests and prescriptions. Journal of Clinical Epidemiology. 2011;64(4):431–5.

[Valkhoff2014] Valkhoff VE, Coloma PM, Masclee GMC, Gini R, Innocenti F, Lapi F, et al. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. Journal of Clinical Epidemiology. 2014

[Vlug1999] Vlug AE, van der Lei J, Mosseveld BM, van Wijk MA, van der Linden PD, Sturkenboom MC, et al. Postmarketing surveillance based on electronic patient records: the IPCI project. Methods Inf Med. 1999;38(4-5):339–44.

[Walley1997] Walley T, Mantgani A. The UK General Practice Research Database. The Lancet. Oct 1997;350(9084):1097–9.

[Williamson2014] Williamson T, Green ME, Birtwhistle R, Khan S, Garies S, Wong ST, et al. Validating the 8 CPCSSN Case Definitions for Chronic Disease Surveillance in a Primary Care Database of Electronic Health Records. Ann Fam Med. 2014;12(4):367–72.

CHAPTER 3

Identifying type 2 diabetes, hypertension and ischaemic heart disease from data sources with incomplete diagnostic information: a population-based validation study in Italian Administrative Databases

Rosa Gini [1,2], Martijn Schuemie [3,4], Alessandro Pasqua [5], Patrizio Dazzi [6], Emanuele Carlini [6], Massimo Coppola [6], Iacopo Cricelli [7], Valentina Barletta [1], Paolo Francesconi [1], Francesco Profili [1], Francesco Lapi [5], Kaatje Bollaerts [8], Andrea Donatini [9], Mario Saugo [10], Mariadonata Bellentani [11], Niek Klazinga [12], Miriam Sturkenboom [2]

1. Agenzia regionale di sanità della Toscana, Osservatorio di epidemiologia; 50141 Florence, Italy
2. Department of Medical Informatics, Erasmus Medical Center; 3015 GJ Rotterdam, The Netherlands
3. Janssen Research & Development, Epidemiology; Titusville, New Jersey, United States
4. Observational Health Data Sciences and Informatics (OHDSI); New York, New York, United States
5. Health Search, Italian College of General Practitioners and Primary Care; 50100 Florence, Italy
6. Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy
7. Genomedics; 50100 Florence, Italy
8. P-95, 3001 Heverlee, Belgium
9. Assessorato Politiche per la Salute, Viale Aldo Moro 21, 40127 Bologna, Italy.
10. Sistema Epidemiologico Regionale Regione Veneto, Padua, Italy;
11. Agenzia Nazionale per i Servizi Sanitari Regionali; 00100 Rome, Italy
12. Academic Medical Center, University of Amsterdam; 1100 DD Amsterdam, The Netherlands

ABSTRACT

Background

Italy has 60 million inhabitants and a universal coverage health care system. The country is divided in around 100 Local Health Units (LHUs), and each LHU collects administrative data about health care delivered to the inhabitants of its community. The data model of Italian Administrative Databases (IAD) is common across the country: information is collected for instance on inpatient care, dispensed drugs, delivery of diagnostic exams and secondary care. Attendance and costs of health care are recorded but not the diagnostic codes in outpatient specialist care, while encounters in primary care are not recorded. In the past years, several algorithms have been proposed to identify patients with chronic conditions in IAD utilizing data from discharge diagnoses, drug utilization or utilization of other healthcare services. This study was initiated to validate the case-finding algorithms for type 2 diabetes mellitus (T2DM), hypertension and ischaemic heart disease (IHD) from the IAD against the case identification in primary care medical records that do have the diagnoses for these conditions.

Methods

All the subjects registered with 25 General Practitioners (GPs) with good quality clinical medical records, and 16 years or older, living in 5 different regions entered the study. Medical records of study subjects were collected from their GPs. IAD data of the same subjects was collected as well from the local health units. Sensitivity, specificity, positive (PPV) and negative (NPV) predictive values of a set of set of case-finding algorithms on IAD were estimated using the medical records as gold standard. The effect on validity indices of adding more years of look-back to the case-finding algorithms was estimated.

Results

33,949 persons entered the study. According to the medical records 2,852 (8.4%) had T2DM, 11,320 (33.2%) had hypertension and 1,414 (4.1%) had IHD.

The algorithms that were used on the IAD records with best balance had sensitivity, specificity, PPV and NPV of, respectively, 72%, 100%, 95%, 98% for T2DM, 68%, 95%, 88%, 86% for hypertension and 44%, 100%, 81%, 98% for IHD. Adding years of look-back improved sensitivity, in particular for IHD. When

drug utilization in less recent years was added as an inclusion criterion, PPV was reduced.

Conclusion
For the three conditions that we investigated, case-finding algorithms on IAD records had excellent specificity and good PPV and NPV, but sensitivity lower than 75% and, in the case of IHD, lower than 50%. Longer look-backs of data are expected to improve these figures, but caution should be adopted in including too many years of look-back from drug utilization. Calibration can be used to assess the impact of imperfect case-finding algorithms on study results.

INTRODUCTION

Italy's population is among the oldest in the world, and good management of chronic conditions, such as diabetes or cardiovascular diseases, is essential to prevent complications and disability, to avoid hospitalization and intense healthcare utilization, and ultimately to ensure sustainability of a universal coverage health care system [OECD2015].

Italy has a tax-based, universal coverage national health system [LoScalzo2009]. Regional health care systems adopt different policies, and the Italian National Agency for Regional Health Systems is mandated to compare quality of healthcare across regions. The MATRICE Project was an initiative of the Agency, aimed to create tools to perform surveillance of prevalence and monitor quality of healthcare for type 2 diabetes (T2DM), hypertension and ischaemic heart disease (IHD) on the basis of Italian administrative databases (IAD), a data source covering the entire national population that has a standardized content.

Many high prevalent chronic conditions are usually diagnosed and taken care of in primary care settings. Unlike other countries such as the United States, in Italy administrative databases do not collect the diagnostic codes recorded during primary care. Each Italian adult inhabitant has the right of choosing a GP, and GPs are paid by LHUs on a capitation fee. Therefore there is no administrative need to record a specific visit, let alone the disease that led to it. Specialist physicians are employees of hospitals, and while the occurrence of an encounter with a patient is recorded in IAD, diagnostic codes are not. Lack of diagnostic codes poses challenges to use these data for quality monitoring as the validity of case-finding algorithms may be affected. For instance, a recent Canadian study found that inpatient data alone underestimate the population prevalence of heart failure by at least 33% [Blais2014]. Although IAD do not have outpatient diagnoses, they do record inpatient diagnosis, as well as utilization of other health care services, such as dispensing of drugs or administration of tests for diagnosis or follow-up. These services may be used as proxies. Moreover, patients with specific diseases are exempted from co-payment to healthcare, and IAD contains a registry of such exemptions. This wealth of data is available for research. In many previous studies using IAD

as a data source, different case-finding algorithms have been used to identify populations with chronic diseases [Maio 2005, Gnavi 2008, Simonato 2008, Chini 2011, Belleudi 2012, Gnavi 2008b, Gnavi 2011, Giorda 2012, Gini 2013b, Buja 2013, Visca 2013, Gini 2014]. However, a formal validation study has never been performed.

In order to provide guidance about using IAD for surveillance and monitoring of quality of healthcare for chronic diseases, the MATRICE Project aimed to conduct a population-based validation study to estimate validity indices of a list of case-finding algorithms detecting T2DM, hypertension and IHD from IAD.

METHODS

Study design

This is a population-based validation study. Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of a list of algorithms detecting T2DM, hypertension and IHD from IAD are estimated in a large sample of the Italian adult population. Primary care medical records are used as a gold standard. The impact on validity of adding one year of look-back to IAD was assessed.

Italian administrative databases

Each Italian region is divided in geographic subareas (on average 10 per region). Healthcare for the population in each area is managed by organizations called Local Health Units (LHU). LHUs collect administrative data on the healthcare they provide to their inhabitants which together form the basis of the IAD.
The main components of IAD are the following tables

- Inhabitant registry (PERSON) is the list of subjects who live in a defined geographical area, recorded with gender, date of birth, date of entry, date of exit, identifier of the chosen GP, citizenship, residence municipality; the dataset is longitudinal, meaning that if a person changes GP, or citizenship, or municipality, a new record is added.
- Hospital discharge records (HOSP) is the table of hospital discharge records reimbursed by the healthcare system, recorded with up to six diagnosis codes and up to six procedure codes in ICD9CM

- Exemption registry (EXE) is the table of disease-specific exemptions from co-payment to the healthcare system, recorded with a disease code which is a truncated ICD9CM code
- Drug dispensing registry (DRUGS) is the table of drugs dispensed by community or hospital pharmacies free of charge or upon co-payment. Drugs are coded with a specific Italian coding system, which is mapped to the Anatomical Therapeutic Chemical classification system (ATC) and to the Defined Daily Dose (DDD) [WHO2016];
- Outpatient services registry (OUTPAT) is a list of outpatient activities dispensed by the healthcare system free of charge or upon co-payment, among which specialist encounters (with no diagnostic code), laboratory or instrumental or bio-imaging diagnostic tests (without results), recorded with a specific Italian coding system

Within a LHU, all the tables above can be linked with each other at the individual level, using the national fiscal identifier as a common key.

Collection of IAD tables is mandatory by national law. The first table to be established was HOSP. LHUs have started to collect IAD tables in different time periods. At 1st January 2012, at least 4 years of look-back for hospital discharge records and 2 of drug dispensing registry were available in all LHUs. At least 3 years of exemption registry were available in 4 LHUs and one year was available in the other. The outpatient services registry was not available in one LHU. The hospital discharge records had a maximum of 11 years in one LHU. The exemption registry had a maximum of 12 years in 3 LHUs. The drug dispensing registry had a maximum of 6 years in one LHU. The outpatient services registry had a maximum of 6 years in one LHU (Table 1).

Primary care medical records

The National College for General Practitioners (SIMG) is the national scientific society of General Practitioners (GPs) in Italy. SIMG has trained the GPs to improve the quality of recording in their medical records. More than 900 members of SIMG use the same clinical software and share their de-identified medical records in Health Search, a database which is regularly used for epidemiological, public health and health services research [Mazzaglia2009, Sultana2014, Gini2014, Gini2016].

Study population

Five regions belonging to the three macro-areas of Italy participated in the MATRICE Project, with one LHU per region: three from the Northern area (LHU 1 from Brescia in Lombardy, LHU 2 from Vicenza in Veneto, LHU 3 from Bologna in Emilia Romagna), one from the Central area (LHU 4 from Arezzo in Tuscany) and one from the Southern area (LHU 5 from Taranto in Puglia). Every Italian inhabitant is entitled to choose a GP, although parents might opt for a specialist paediatrician instead for their children, up to the age of 15. Therefore, each inhabitant older than 15 is registered with and in charge of a dedicated GP. SIMG identified, in each of the participating LHUs, five GPs that participated also in the Health Search database at January 1st 2012. The study population was composed of the persons registered with these 25 GPs on January 1st 2012, and older than 15 at the same date.

Clinical definitions of the diseases

A panel of cardiologists, diabetologists, epidemiologists and experts in organization of primary care services participating in the MATRICE Project established clinical definitions of T2DM, hypertension and IHD, which are shown in Table 2 of Chapter 2 of this thesis.

Selection of algorithms

Using the clinical definitions in Supplementary Table 1 as a reference, a panel of experts selected ICD9CM codes corresponding to diagnoses of the three conditions, ATC codes of the drugs indicated for treatment, and codes for the most common follow-up exams. Italian literature and gray literature was searched to obtain algorithms that had commonly been used. A national workshop gathered the main Italian investigators in the field to validate the final list of candidate algorithms on IAD to be tested in this study.

Component and composite algorithms

To streamline data processing and analysis, the algorithms from the final list were divided in components. Each component required data from a single data table.

The complete list of components can be found as Supplementary Material (www.ars.toscana.it/files/progetti/informatica_medica/thematrix/SupplementaryMaterial.zip). Each component algorithm was described as a sequence of two steps:

1. record selection from a single table among HOSP, EXE, DRUGS and OUTPAT;
2. identification of subjects with a specific pattern of records in the selection (for instance: at least two in a 365 days).

To obtain composite algorithms, components were combined by means of logical operators (OR, AND, AND NOT).

Gold standard

Case-finding algorithms on primary care medical records for the three conditions were tested in a previous study against a questionnaire submitted to the GPs, to assess their validity [Gini2015]. Algorithms had excellent PPV for T2DM and hypertension (100%), and very high for IHD (98%), therefore the number of false positives is expected to be negligible. The number of false negatives is expected to be very low as well among diagnosed cases, because the prevalence of the diseases we observed in this study are higher than the estimates obtained in other data sources, such as national surveys [Gini2013,Cricelli2003].
We assumed that medical records were providing a perfect description of diagnosed cases.

Data collection

All administrative records from the main IAD tables collected by the 5 LHUs on the study population were sent to the National Council of Research (CNR). The medical records of the 25 GPs were queried using the gold standard algorithms, and the resulting datasets were sent to CNR as well.
Before transmission to CNR, fiscal codes were pseudonymised from both types of data sources using the same encryption key, which was transmitted to LHUs and GPs by the Agenzia Regionale di Sanità della Toscana (ARS). Due to the use of the same encryption key, CNR could perform deterministic record linkage between the medical records arriving from GPs and the IAD data from the LHU. Data extraction, pseudonymization and transmission were performed automatically by a tailored suite of software tools.

Data processing

Each component case-finding algorithm was applied to the linked dataset. In a first data processing step, a look-back period homogeneous across LHUs was used: records from hospitalizations were queried for 4 years (from 1st January 2008 to 31st December 2011), exemptions for 3 years (from 1st January 2009 to 31st December 2011), while drugs and outpatient visit records were queried for 2 years, from 1st January 2010 to 31st December 2011, where data was available. In the second data processing step, all the algorithms were applied repeatedly, adding one year of look-back per data table at a time, where data was available. Data processing was embedded in the domain-specific language of an open-source software tool developed for MATRICE [TheMatrix].

The final dataset contained a single record per subject, with gender, age band, GP, LHU, and, for each disease, the values of case-finding algorithms and of the gold standard. The pseudonymized subject identifiers were removed and the dataset was transmitted to ARS for data analysis. Figure 1 is a graphical representation of this process.

Data analysis

The prevalence of diagnosed T2DM, hypertension and IHD in the study population was estimated from the gold standard.

For each case-finding algorithm, sensitivity, specificity, PPV and NPV were computed, for the overall population and stratifying by LHU. Sensitivity was the proportion of subjects positive for the gold standard that were positive for the algorithm; specificity was the proportion of subjects negative for the gold standard that were negative for the algorithm; PPV was the proportion of subjects positive for the algorithm that were positive for the gold standard; NPV was the proportion of subjects negative for the algorithm that were negative for the gold standard.

For a selection of algorithms (both in terms of significance and of performance in the crude analysis) adjusted sensitivity, specificity, PPV and NPV were estimated. We adjusted for the fact that these metrics might depend on gender, age, and LHU, and GPs have different distributions of these dependent variables. This was done by averaging predictive margins per GP in a random effects model [Williams2012]. Heterogeneity of sensitivity, specificity, PPV and NPV across sites was estimated with the Wald test of significance for the variable LHU in the corresponding model.

The algorithm with best balance among sensitivity and PPV ('recommended algorithm') was identified per each disease.
In some LHUs more years of look-back were available. If there is a longer look back period for a person there is more opportunity to have a diagnosis or some disease proxy, and that would increase sensitivity, maybe at the expense of a lower PPV. We first computed the validity indices of the recommended algorithms using all available data. We created for each disease a dataset including, per each combination of available look-back years of hospital discharge records, exemption and drug dispensing registry, all the study subjects from those LHUs which had the corresponding span of look-back available. For simplicity, we truncated the available years to 6 or more, so the triples ranged from (4,3,2) to (6+,6+,6+). In each copy we computed the algorithm for all the study subjects using the corresponding triple of look-back years. We then estimated sensitivity and PPV from LHU-specific logistic models, using the triple of numbers of look-back years as independent variables with interactions, adjusted by gender and age. We finally estimated, per triple, the variation of sensitivity and PPV when adding a new year of look-back of each data table as the mean value among LHU-specific variations.
Statistical analysis was performed using Stata version 12.1 (Stata Corporation).

Ethics
Permission to perform record linkage between IAD and medical records was granted by the Italian National Authority for the Privacy regulation. Specifically, permission was granted to CNR to store and process the data, and to ARS to obtain the linked individual-level analytical dataset, for statistical analysis.

RESULTS

Study population
The study population from the LHUs comprised 36,414 subjects, 34,560 (94.9%) of which had complete and unambiguous data (gender, birth date, date of entry and exit) and were present in the LHU and older than 15 years at the index date. The 25 GPs sent data on 34,933 subjects. Data on 33,949 persons (98.2% of those from LHUs, 97.2% of those from GPs) could be linked across the two data sources. Of this study population 51.7% was composed of female persons, 39.2%

was 16-44 years old, 33.6% was 45-64 years old, 22.9% was 65-84 years old and 4.3% was 85 or older. LHU 3 had the largest share of population, and LHU 4 the smallest. The population in LHU 5 was slightly younger. According to the gold standard, 2,852 (8.4%) persons had T2DM, 11,320 (33.2%) had hypertension and 1,414 (4.1%) had IHD in the study population (Table 1).

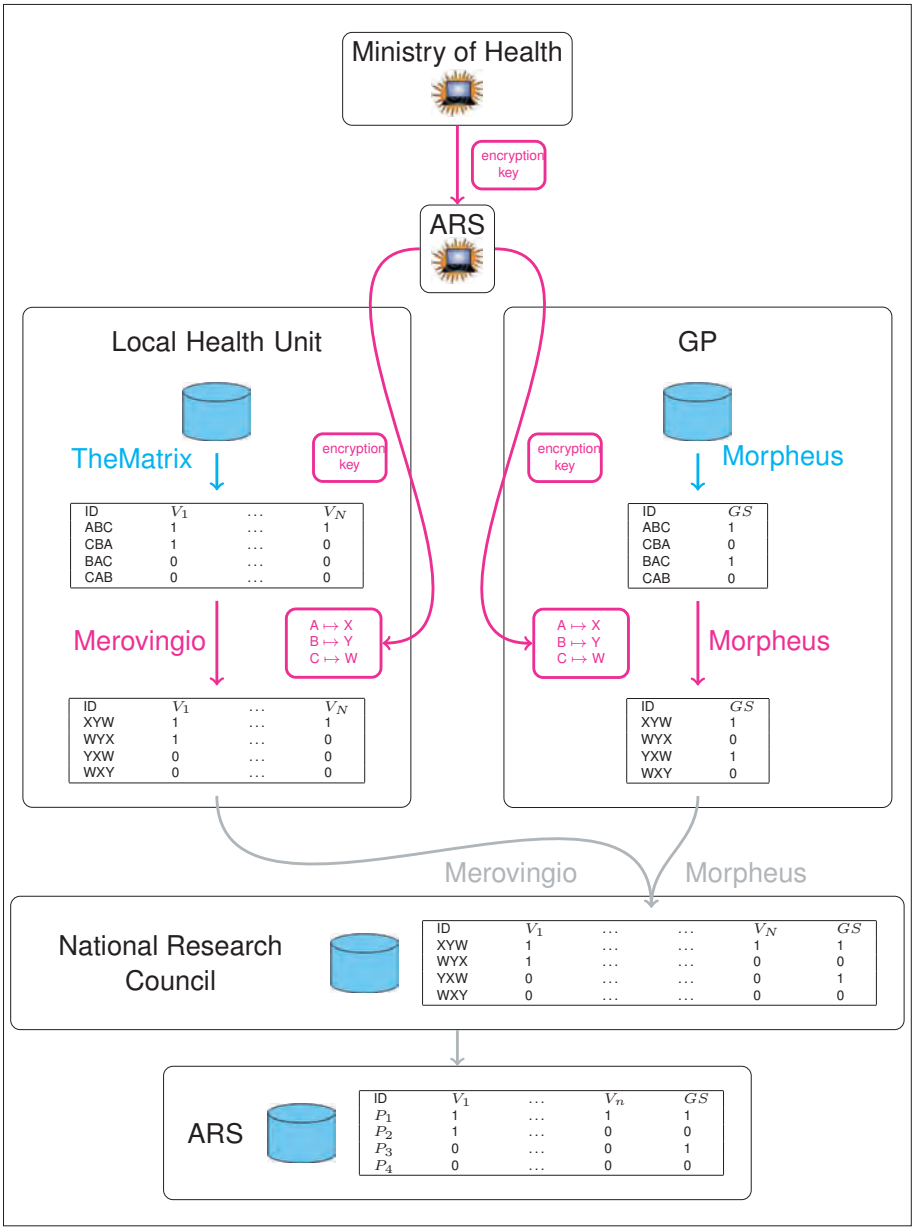
Table 1. Study population and years of look-back per LHU. HOSP: hospital discharge records, EXE: exemptions rgistry; DRUGS: drug dispensing registry. OUTPAT: outpatient activity registry.

		LHU 1	LHU 2	LHU 3	LHU 4	LHU 5	Total
N		6,949	6,803	8,040	5,756	6,401	33,949
F		3443 (49.5)	3499 (51.4)	4306 (53.6)	3017 (52.4)	3281 (51.3)	17546 (51.7)
Age	16-44	2835 (40.8)	2605 (38.3)	2916 (36.3)	2130 (37.0)	2815 (44.0)	13301 (39.2)
	45-64	2514 (36.2)	2246 (33.0)	2624 (32.6)	1813 (31.5)	2201 (34.4)	11398 (33.6)
	65-84	1424 (20.5)	1643 (24.2)	2082 (25.9)	1456 (25.3)	1171 (18.3)	7776 (22.9)
	85+	176 (2.5)	309 (4.5)	418 (5.2)	357 (6.2)	214 (3.3)	1474 (4.3)
GP	1	1464 (21.1)	1526 (22.4)	1755 (21.8)	1416 (24.6)	1416 (22.1)	7577 (22.3)
	2	1444 (20.8)	1509 (22.2)	1682 (20.9)	1156 (20.1)	1372 (21.4)	7163 (21.1)
	3	1407 (20.2)	1492 (21.9)	1664 (20.7)	1137 (19.8)	1323 (20.7)	7023 (20.7)
	4	1373 (19.8)	1262 (18.6)	1620 (20.1)	1086 (18.9)	1288 (20.1)	6629 (19.5)
	5	1261 (18.1)	1014 (14.9)	1319 (16.4)	961 (16.7)	1002 (15.7)	5557 (16.4)
Disease	T2DM	567 (8.1)	580 (8.5)	594 (7.4)	478 (8.2)	633 (9.9)	2,852 (8.4)
	Hypertension	2,456 (35.4)	2,340 (34.0)	2,701 (33.4)	1,799 (31.5)	2,024 (31.7)	11,320 (33.2)
	IHD	258 (3.7)	275 (4.0)	433 (5.4)	235 (4.0)	213 (3.4)	1,414 (4.1)
Years of look-back	HOSP	11	7	6	6	4	
	EXE	12	12	12	3	1	
	DRUGS	3	5	2	6	2	
	OUTPAT	3	3	2	6	0	

Validity indices

The validity indices of the list of algorithms are shown in Supplementary Table 3. Table 2 shows the validity indices of the most relevant algorithms, which are represented also in Figure 2, for T2DM, hypertension and IHD. In all three cases, algorithm A is the one detecting subjects only from discharge diagnoses: the PPV is very high (from 84% for IHD to 94% for T2DM), but the sensitivity is lower than 25% (from 22% of IHD to 11% of hypertension). Adding subjects with a disease-specific exemption (algorithm B) does not change the PPV but improves the sensitivity slightly, up to 30% in both T2DM and IHD. Adding persons utilizing specific drugs (algorithm C) has a different impact on the three conditions: in

Figure 1 Data collection process



the case of T2DM inclusion of non-insulin glucose lowering drugs increases sensitivity dramatically to 67% while the PPV remains the same. In the case of hypertension utilization of selective antihypertensive drugs (alpha blockers) has only a modest impact on the indices of the algorithm. For IHD utilization of nitrates increases sensitivity to 44% while significantly lowering PPV to 80%. For IHD algorithms D and E (respectively, adding subjects who utilize platelet aggregation inhibitors excluding heparin; and adding only those subjects who utilize both this class of drugs *and* beta-blockers) improves the sensitivity greatly (respectively, to 82% and 69%) but at the price of lowering dramatically the PPV (29% and 51% respectively). In the case of hypertension, inclusion of beta-blockers or renin-angiotensin agents (algorithm D) leaves the PPV almost unchanged (88%) while the sensitivity (68%) is increasing substantially. In the case of T2DM addition of persons with repeated measurements of glycated haemoglobin (a component algorithm which is only available in 4 of the 5 LHUs) increased the sensitivity to 75%, but the PPV was reduced (90%); in algorithm E insulin use is added, which increases the sensitivity to 71% while keeping PPV at 94%.

The best balance between PPV and sensitivity is therefore Algorithm E for T2DM, algorithm D for hypertension and Algorithm C for IHD. They entered the following analysis as 'recommended algorithms'.

Variation in positive predictive value between LHUs is not significant for diagnostic components in T2DM and IHD, but it is significant for hypertension; for drug utilization components it is significant for T2DM and hypertension (in both cases LHU 4 had worst performance), but not for IHD. The variability in sensitivity of the preferred algorithm across LHUs is high for hypertension, with LHU 5 showing worst performance: this is because the drugs component adds less. Heterogeneity in sensitivity is important for IHD as well, due to high contribution from exemptions in LHU 1 (Table 2 and Supplementary Table 3).

Adding more years of look-back

When we pooled all available look-back years, the crude estimates of sensitivity, specificity, PPV and NPV of the recommended algorithms were, respectively: 76%, 99%, 86%, 98% for T2DM, 73%, 93%, 83% e 87% for hypertension and 63%, 98%, 79% e 99% for IHD.

The impact of adding years in the look-back period for hospitalizations and exemptions was modest in the case of T2DM and hypertension, but very

Figure 2. Adjusted sensitivity and PPV of the main algorithms for (from left to right) type 2 diabetes, hypertension and ischaemic heart disease. The orange square contains algorithms with both indices higher than 60%, the green square contains algorithms with both indices higher than 75%. Algorithm A: subjects with an inpatient diagnosis; algorithm B: subjects in A OR with an exemption; algorithm C: subjects in B OR in drug utilization (T2DM: oral antidiabetics; HYPERTE: antihypertensives; IHD: nitrates); algorithm D: subjects in C OR other (T2DM: repeated glycated hemoglobin measurements; HYPERTE: beta-blockers or renin-angiotensin agents: IHD: platelet aggregation inhibitors excluding heparin); E: subjects in C OR other (T2DM: insulin utilization; IHD: platelet aggregation inhibitors excluding heparin AND betablockers)

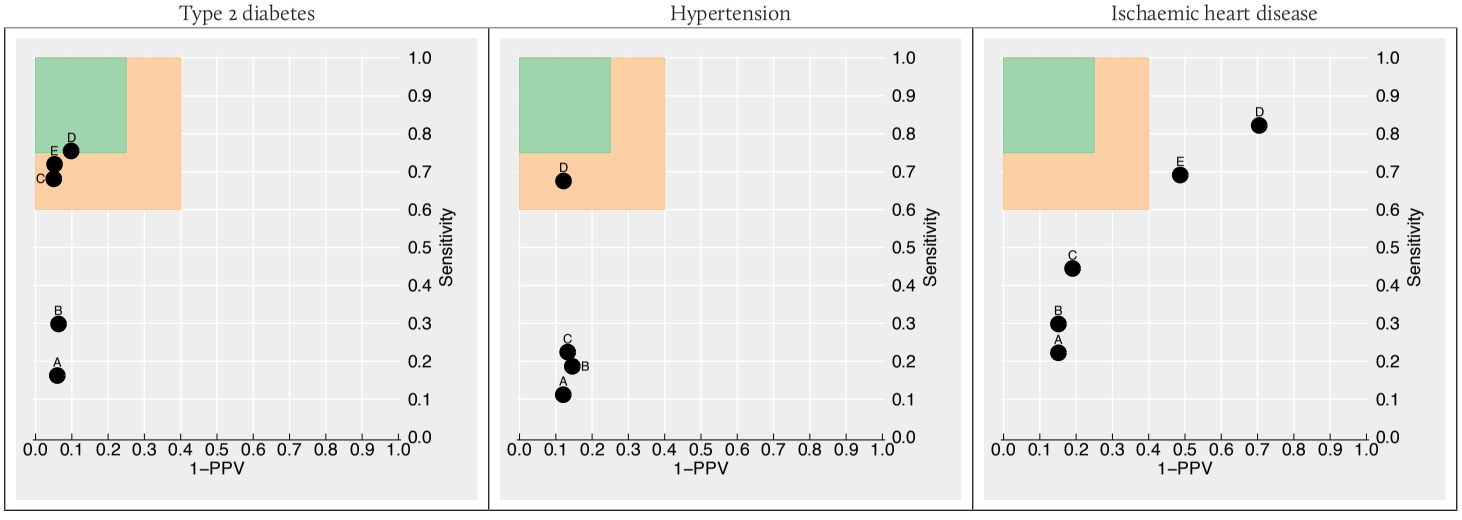


Table 2. Validation indices of the main algorithms. Algorithm A: subjects with an inpatient diagnosis; algorithm B: subjects in A OR with an exemption; algorithm C: subjects in B OR in drug utilization (T2DM: non-insulin antidiabetics; HYPERTE: antihypertensives; IHD: nitrates); algorithm D: subjects in C OR other (T2DM: repeated glycated hemoglobin measurements; HYPERTE: betablockers or renin-angiotensin agents: IHD: platelet aggregation inhibitors excluding heparin); E: subjects in C OR other (T2DM: insulin utilization; IHD: platelet aggregation inhibitors excluding heparin AND beta-blockers). For each algorithm each validity index is computed from the pooled dataset (crude, pooled), from the dataset stratified for LHU (crude, range), from a logistic model with random effect (estimate and CI%), having GP as a panel variable. The *p* value is the Wald test for the LHU variable in the same model, testing whether the index is significantly different across LHUs. In green: recommended algorithms.

Algorithm		Type 2 diabetes					Hypertension					Ischaemic heart disease				
		Crude		Adjusted			Crude		Adjusted			Crude		Adjusted		
	Index	Pooled	Range	Estimate	CI 95%	p	Pooled	Range	Estimate	CI 95%	p	Pooled	Range	Estimate	CI 95%	P
A	Sensitivity	0.17	(0.06-0.26)	0.16	(0.14-0.19)	<.001	0.11	(0.02-0.19)	0.11	(0.10-0.13)	<.001	0.22	(0.12-0.30)	0.22	(0.19-0.25)	<.001
	Specificity	1.00	(1.00-1.00)	1.00	(1.00-1.00)	0.195	0.99	(0.99-1.00)	0.99	(0.99-0.99)	<.001	1.00	(1.00-1.00)	1.00	(1.00-1.00)	0.54
	PPV	0.94	(0.91-0.97)	0.94	(0.92-0.96)	0.407	0.87	(0.83-0.96)	0.88	(0.86-0.90)	<.05	0.84	(0.73-0.97)	0.85	(0.80-0.90)	0.111
	NPV	0.93	(0.92-0.94)	0.93	(0.92-0.93)	<.05	0.69	(0.66-0.72)	0.69	(0.67-0.71)	<.05	0.97	(0.96-0.97)	0.97	(0.96-0.97)	<.05
B	Sensitivity	0.30	(0.20-0.52)	0.30	(0.26-0.34)	<.001	0.19	(0.08-0.36)	0.19	(0.17-0.21)	<.001	0.30	(0.19-0.48)	0.30	(0.27-0.33)	<.001
	Specificity	1.00	(1.00-1.00)	1.00	(1.00-1.00)	0.854	0.98	(0.94-1.00)	0.98	(0.98-0.99)	<.001	1.00	(0.99-1.00)	1.00	(1.00-1.00)	<.05
	PPV	0.93	(0.90-0.96)	0.94	(0.91-0.96)	0.651	0.85	(0.74-0.97)	0.85	(0.82-0.89)	<.001	0.84	(0.76-0.94)	0.85	(0.81-0.89)	0.56
	NPV	0.94	(0.93-0.96)	0.94	(0.93-0.95)	<.001	0.71	(0.67-0.76)	0.71	(0.69-0.73)	<.05	0.97	(0.96-0.98)	0.97	(0.97-0.97)	<.001
C	Sensitivity	0.67	(0.65-0.69)	0.68	(0.65-0.72)	0.817	0.22	(0.14-0.38)	0.22	(0.20-0.25)	<.001	0.44	(0.36-0.60)	0.44	(0.41-0.48)	<.001
	Specificity	1.00	(0.99-1.00)	1.00	(1.00-1.00)	0.406	0.98	(0.94-1.00)	0.98	(0.98-0.98)	<.001	1.00	(0.99-1.00)	1.00	(0.99-1.00)	0.65
	PPV	0.94	(0.92-0.97)	0.95	(0.93-0.96)	0.121	0.86	(0.74-0.97)	0.87	(0.84-0.89)	<.001	0.80	(0.71-0.87)	0.81	(0.77-0.85)	0.140
	NPV	0.97	(0.96-0.97)	0.97	(0.97-0.98)	<.05	0.72	(0.69-0.77)	0.72	(0.70-0.74)	<.05	0.98	(0.96-0.98)	0.98	(0.97-0.98)	<.001
D	Sensitivity	0.75	(0.65-0.80)	0.75	(0.73-0.78)	<.05	0.67	(0.49-0.80)	0.68	(0.65-0.70)	<.001	0.82	(0.74-0.85)	0.82	(0.80-0.85)	0.178
	Specificity	0.99	(0.98-1.00)	0.99	(0.99-0.99)	<.001	0.95	(0.90-0.98)	0.95	(0.95-0.96)	<.001	0.91	(0.87-0.95)	0.92	(0.91-0.92)	<.001
	PPV	0.90	(0.80-0.97)	0.90	(0.88-0.92)	<.001	0.87	(0.78-0.94)	0.88	(0.86-0.90)	<.001	0.29	(0.22-0.35)	0.30	(0.27-0.32)	<.05
	NPV	0.98	(0.96-0.98)	0.98	(0.97-0.98)	<.001	0.85	(0.80-0.91)	0.86	(0.84-0.87)	<.001	0.99	(0.99-0.99)	0.99	(0.99-0.99)	0.353
E	Sensitivity	0.71	(0.68-0.76)	0.72	(0.69-0.75)	0.516						0.69	(0.62-0.76)	0.69	(0.67-0.72)	<.05
	Specificity	1.00	(0.99-1.00)	1.00	(1.00-1.00)	0.527						0.97	(0.95-0.98)	0.97	(0.97-0.97)	<.001
	PPV	0.94	(0.92-0.97)	0.95	(0.93-0.96)	0.97						0.51	(0.39-0.60)	0.51	(0.49-0.54)	<.001
	NPV	0.97	(0.97-0.98)	0.98	(0.97-0.98)	<.05						0.99	(0.98-0.99)	0.99	(0.98-0.99)	0.216

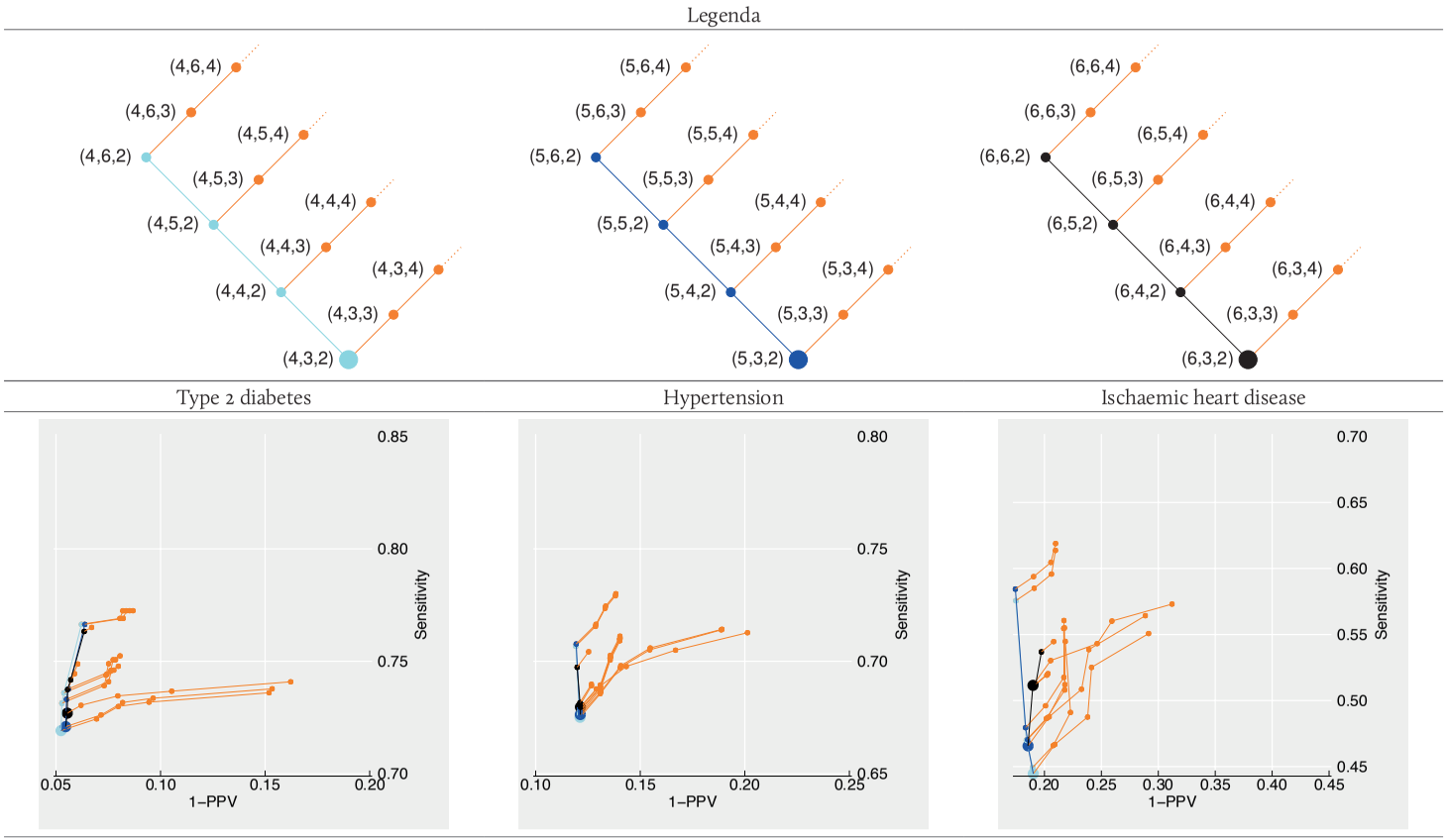
Table 3. Average variation in sensitivity and PPV across LHUs, when adding one further year of look-back, in percentage points. HOSP: hospital discharge records, EXE: exemptions registry; DRUGS: drug dispensing registry. OUTPAT: outpatient activity registry. In italics the figures that refer to the variation when adding one year *or more* (from 5 years on).

Years of look -back			Average variation when adding one further year of look-back, in percentage points																	
HOSP	EXE	DRUGS	Type 2 diabetes						Hypertension						Ischaemic heart disease					
			HOSP		EXE		DRUGS		HOSP		EXE		DRUGS		HOSP		EXE		DRUGS	
			SE	PPV	SE	PPV	SE	PPV	SE	PPV	SE	PPV	SE	PPV	SE	PPV	SE	PPV	SE	PPV
4	3	2	0.2	-0.2	1.2	0.0	0.5	-1.7	0.1	0.0	0.2	0.0	1.1	-1.0	2.1	0.5	0.5	0.1	2.2	-1.9
		3	0.2	-0.2	1.7	0.1	0.6	-1.0	0.1	0.0	0.1	0.0	1.2	-1.0	2.2	0.6	0.5	0.2	2.1	-2.9
		4	0.1	-0.1	1.4	0.1	0.2	-1.5	0.0	0.0	0.1	0.0	0.8	-1.4	2.5	0.8	0.7	0.3	3.8	-0.3
		5	0.1	-0.1	1.4	0.1	0.4	-5.7	0.0	0.0	0.1	0.0	0.9	-3.4	1.8	0.5	0.7	0.2	2.6	-5.0
		6	0.0	-0.2					0.0	0.0					1.7	0.4				
	4	2	0.2	-0.2	0.5	-0.1	0.8	-2.0	0.1	0.0	0.1	0.0	1.1	-1.0	2.1	0.5	2.1	0.6	1.7	-1.8
		3	0.4	-0.1	0.5	-0.1	0.5	-0.1	0.1	0.0	0.1	0.0	1.3	-0.5	2.2	0.7	2.5	0.8	2.5	-1.5
		4	0.2	0.0	0.5	0.1	0.2	-0.3	0.0	0.0	0.1	0.0	0.9	-0.5	2.9	1.0	4.7	1.5	5.4	0.5
		5	0.2	0.0	0.5	0.1			0.0	0.0	0.1	0.0			1.8	0.5	3.9	1.2		
	5	2	0.2	-0.2	*3.0*	*-0.8*	0.8	-2.0	0.1	0.0	*2.9*	*0.2*	1.0	-1.0	0.9	0.1	*10.6*	*0.9*	1.7	-1.8
		3	0.4	-0.1	*2.9*	*-0.9*	0.5	-0.1	0.1	0.0	*3.3*	*0.2*	1.3	-0.5	1.0	0.3	*12.4*	*2.1*	2.2	-1.6
		4	0.2	0.0	*2.9*	*-0.7*	0.2	-0.3	0.0	0.0	*4.3*	*0.3*	0.9	-0.5	0.4	0.1	*23.7*	*4.3*	4.7	0.1
		5	0.2	0.0	*2.7*	*-0.7*			0.0	0.0	*4.0*	*0.3*			0.0	0.0	*20.8*	*3.8*		
	6+	2	0.1	-0.1			0.3	-1.8	0.1	0.0			0.9	-0.9	0.9	0.1			0.9	-1.6
		3	0.1	-0.1			0.3	-0.1	0.0	0.0			0.8	-0.5	1.0	0.2			1.1	-1.5
		4	0.0	0.0			0.0	-0.3	0.0	0.0			0.6	-0.5	0.4	0.1			1.8	-0.3
		5	0.0	0.0			.	.	0.0	0.0					0.0	0.0				
5	3	2	*0.6*	*-0.1*	1.2	0.0	0.5	-1.7	*0.3*	*0.0*	0.2	0.0	1.0	-1.0	*4.6*	*-0.4*	0.5	0.0	2.2	-1.9
		3	*0.8*	*-0.1*	1.7	0.1	0.6	-1.0	*0.4*	*0.0*	0.1	0.0	1.2	-1.0	*6.0*	*-0.5*	0.5	0.2	2.1	-2.8
		4	*0.2*	*0.1*	1.4	0.1	0.2	-1.5	*0.1*	*0.0*	0.1	0.0	0.8	-1.4	*3.4*	*0.4*	0.7	0.2	3.0	-0.6
		5	*0.2*	*0.1*	1.4	0.1	*0.4*	*-5.7*	*0.1*	*0.0*	0.1	0.0	*0.8*	*-3.4*	*3.4*	*0.5*	0.7	0.2	*2.6*	*-5.0*
		6	*0.2*	*0.1*					*0.0*	*0.0*					*2.1*	*0.1*				

Table 3. Continued

Years of look -back			Average variation when adding one further year of look-back, in percentage points																	
HOSP	EXE	DRUGS	Type 2 diabetes						Hypertension						Ischaemic heart disease					
			HOSP		EXE		DRUGS		HOSP		EXE		DRUGS		HOSP		EXE		DRUGS	
			SE	PPV	SE	PPV	SE	PPV	SE	PPV	SE	PPV	SE	PPV	SE	PPV	SE	PPV	SE	PPV
5	4	2	*0.8*	*-0.1*	0.5	-0.1	0.8	-2.0	*0.5*	*0.0*	0.1	0.0	1.1	-1.0	*5.2*	*-0.7*	0.9	0.2	1.7	-1.8
		3	*1.4*	*-0.3*	0.5	-0.1	0.5	-0.1	*0.7*	*0.0*	0.1	0.0	1.3	-0.5	*8.6*	*-1.2*	1.3	0.4	2.5	-1.5
		4			0.5	0.1	0.2	-0.3			0.1	0.0	0.9	-0.5			2.2	0.7	4.3	0.0
		5			0.5	0.1					0.1	0.0					2.2	0.6		
	5	2	*0.8*	*-0.1*	*2.9*	*-0.8*	0.8	-2.0	*0.5*	*0.0*	*2.9*	*0.2*	1.0	-1.0	*5.2*	*-0.7*	*10.5*	*0.9*	1.7	-1.8
		3	*1.4*	*-0.3*	*2.6*	*-1.0*	0.5	-0.1	*0.7*	*0.0*	*3.3*	*0.2*	1.3	-0.5	*8.6*	*-1.2*	*12.4*	*2.1*	2.2	-1.6
		4			*2.7*	*-0.8*	0.2	-0.3			*4.3*	*0.3*	0.9	-0.5			*23.7*	*4.2*	4.3	0.0
		5			*2.6*	*-0.8*					*4.0*	*0.3*					*20.8*	*3.8*		
	6+	2	*0.4*	*-0.2*			0.3	-1.8	*0.3*	*0.0*			0.9	-0.9	*4.9*	*-0.6*			0.9	-1.6
		3	*0.5*	*-0.3*			0.3	-0.1	*0.3*	*0.0*			0.8	-0.5	*8.6*	*-1.2*			1.1	-1.5
		4					0.0	-0.3					0.6	-0.5					1.4	-0.4
6+	3	2			1.1	0.0	0.4	-0.7			0.1	0.0	0.8	-0.8			0.0	-0.1	1.9	-1.6
		3			1.9	0.2	0.4	-1.8			0.2	0.0	1.0	-1.4			0.0	0.0	1.3	-4.1
		4					0.2	-2.6					0.7	-2.4					1.7	-1.3
		5					0.4	-5.7					0.8	-3.4					1.3	-5.3
	4	2			0.4	-0.2	0.7	-0.3			0.1	0.0	0.9	-0.5			0.1	-0.1	0.8	-1.1
		3			0.5	-0.2					0.1	0.0					0.0	0.0		
	5	2			*2.2*	*-0.6*	0.7	-0.3			*1.6*	*0.1*	0.8	-0.5			*2.4*	*-0.5*	0.8	-1.1
		3			*1.4*	*-1.2*					*1.4*	*0.1*					*0.0*	*0.0*		
	6+	2					0.2	-0.4					0.7	-0.6					0.8	-1.1

Figure 3. Sensitivity and PPV of the preferred algorithms for (from left to right) type 2 diabetes, hypertension and ischaemic heart disease, when adding years of look-back. HOSP: hospital discharge records, EXE: exemptions registry; DRUGS: drug dispensing registry. OUTPAT: outpatient activity registry. Big markers in shades of blue represent adding HOSP data only, starting from 4 (from lighter to darker shades of blue). The paths connecting small markers in shades of blue represent adding EXE data to HOSP, each step in the path represents one year, starting from 3. The paths connecting small orange markers represent adding DRUGS data to HOSP and/or EXE, each step in the path represents one year, starting from 2. Axis scales are adapted to the graphs.



relevant, especially the latter, in the case of IHD, without loss in PPV. Adding years of look-back to the drug registry produced a small increase in sensitivity, at the expense of the PPV, especially in the case of T2DM (Figure 3 and Table 3).

DISCUSSION

This paper provides the validity of algorithms to identify type 2 diabetes mellitus, hypertension and ischaemic heart disease in Italian administrative databases, which miss primary care diagnostic information.

We demonstrated that exploiting both diagnostic and other available information in IAD, algorithms with very good PPV could be found for T2DM (94%) and hypertension (88%), and with good PPV for IHD (80%). However, sensitivity was suboptimal for T2DM (71%) and hypertension (68%), and low (44%) for IHD. Adding more years of look-back to hospital discharge records and exemption registry is likely to improve substantially sensitivity of the IHD algorithm: we estimated that more than 15 percentage points would be added if more 6 years of look-back in the two data sources were included in the algorithm. Adding more time to the two years of look-back of the drug registry seemed to reduce PPV, particularly in the case of T2DM.

Sensitivity and positive predictive value of diagnoses and other component algorithms in IAD

Diagnostic codes are only collected in Italian Administrative Databases during inpatient care or when a disease-specific exemption from copayment is dispensed. As expected, this hampers the sensitivity of algorithms selecting subjects recorded with diagnostic codes. Persons aged 65 or more are exempt from any copayment, regardless of a disease, therefore patients who develop a chronic condition in older age are not recorded in the exemption registry. Hospital discharge records are more sensitive in older ages, because older patients are more often admitted to hospital, regardless of the condition. Hospital discharge records and exemption registry are therefore complementary, the former being more sensitive in older age bands and the latter in the younger. The consequence is that in T2DM, a condition which arises at younger age with respect to IHD, hospital discharge records are less sensitive, and exemption registry is complementary. Misclassification is low both inpatient setting and in the exemption registry.

The amount of misclassification associated with drug utilization depends on the prevalence of the other conditions which are indications for the same drug. In the case of non-insulin anti-diabetics for T2DM, which have some low-prevalence off-label indications, misclassification is negligible. In the case of platelet aggregation inhibitors for IHD, the prevalence of other indications (atrial fibrillation, ischaemic stroke, transient ischaemic attack, high cardiovascular risk) is very high, and the amount of misclassification is correspondingly high. Even if hypertension is not the only indication for beta-blockers, the prevalence of the other indications is low in comparison, and this explains why this component does not reduce dramatically the PPV of the composite algorithm for hypertension. Sensitivity of drug utilization components is high if treatment with the drug is needed early in the development of the condition, as is the case of non-insulin antidiabetics for T2DM. Moreover, adherence must be high in the population for a drug utilization component to be sensitive: this is not the case of antihypertensives for hypertension in Italy [Poluzzi2005, Corrao2011], and this explains the observed imperfect sensitivity.
Outpatient activity registry was not available in all the LHUs at the index date of our analysis, nevertheless an algorithm based on repeated tests showed a promising performance for T2DM in the LHUs where it was available (PPV of 79% or higher, sensitivity 40% or higher). Similar components are active part of algorithms used in Denmark administrative databases, which share many features of IAD [Carstensen2008]

Adding look-back years
Adding years of hospital discharge records and exemptions had a modest impact in the case of T2DM. Patients who have had a hospital admission or an exemption in a more remote past, due to the natural evolution of the disease, are likely to be now in treatment. They are therefore included in the composite algorithm via the drug utilization component. The opposite is true in the case of IHD: patients with a more remote history of infarction, or with an old exemption, may be non-adherent to their preventive treatment, and would therefore not be detected by the drug utilization algorithm. The contribution of remote years from hospital discharge records and exemption registry becomes therefore relevant. Hypertension is rarely mentioned as a discharge diagnosis, as can be observed from the low sensitivity of algorithm A from Table 2, so it

is unsurprising that adding more years of hospital discharge records does not improve the characteristics of this algorithm.
In all cases, however, long look-backs of discharge records and exemptions proved to be consistent in PPV. An exception is the case of T2DM, where a small decrease in PPV for exemptions may be due to misclassification with type 1 diabetes: indeed, the truncated version of ICD9CM adopted in the exemption registry does not tell one form of diabetes from the other.

Comparison with literature
Multiple algorithms were tested by Morley et al to identify patients with atrial fibrillation from the National Linked Electronic Health Records in the United Kingdom. They assessed the contribution from non-diagnostic algorithms, such as treatment or procedures, to a composite case-finding algorithm, and found that procedures were not a signifiacant component, while treatments could help in establishing more precisely the date of onset of the condition [Morley2014].
In a study with a design similar to ours, Li et al tested various case-finding algorithms for systolic dysfunction in a claims database, using medical records as a gold standard. Similarly, they found low sensitivity and high PPV. They provided examples of bias induced by imperfect validity of their algorithms, and discussed when it resulted in an acceptable error of the resulting estimate. This depended critically on the research question [Li2011].
In a study on febrile convulsions, Quantin et al validated four case-finding algorithms, and found the optimal balance between validity and power to conduct a study on vaccine safey [Quantin2013].
Brunelli et al observed that using all the available look-back time of each person, rather that a fixed window for all the population reduced confounding [Brunelli2013]. Based on our results, caution should be taken in generalizing this finding, because longer look-back time may hamper validity of the variables included in the analysis.

Application of the results of this study
When a study variable is obtained from a case-finding algorithm with known validity indices, it is possible to calibrate the statistical models that include the variable, and adjust the resulting estimates [Green1983, Flegal1986, Magder1997]. For instance, calibration to estimate prevalence for surveillance

purposes consists in multiplying the estimate by the PPV, and dividing it by sensitivity [Rogan1978, Leong2013a].
Quality of health care in cohorts of patients with chronic diseases in Italy is typically measured by means of compliance with standards of care. In a previous study on aggregated Italian data we found that IAD estimated compliance with standards of care for diabetes and IHD consistently with GP data across Italian regions [Gini2014]. Similarly, in a recent study on Medicare data patients false-positive to a case-finding algorithm for diabetes were shown to have a similar healthcare utilization profiles, thus suggesting that the population detected by the algorithm could be used to estimate measures of service utilization [Sakshaug2014]. In general, it is possible that for some specific purposes the populations detected by an imperfect algorithm share the same relevant characteristics with the true target population so that calibration is not necessary.

Recommendations
Overall, adding years of look-back for discharge records and exemption registry, as soon as they become available, is recommended. The same is not true for drug registry.
Beyond adding longer spans of look-back, new components may be added in the near future. Components based on outpatient activity registry look promising. New treatments, more specific for each disease, may become available, and this would provide new components as well. Adherence may improve, and this may increase sensitivity of drug components. To conduct this validation study, a suite of tailored software tools was developed, and permission was granted by the Italian National Authority for the Privacy regulation to perform record-linkage. This experience can now be repeated periodically, when the need arises. This would allow testing new algorithms and update validity indices. This would keep up-to-date the capacity of the Italian investigators to perform studies on T2DM, hypertension and IHD using IAD. In particular, this would maintain confidence in estimates of compliance with standards of care estimated with IAD.
The generalizability of our estimates to the national level can be challenged: heterogeneity across LHUs was high for many indices. For instance, LHU 4 having low PPV for non-insulin antidiabetic drugs utilization may be linked to a frequent off-label use of metformin as a treatment of obesity that is anecdotally

widespread in Tuscany (Fabio Baccetti, personal communication) but not in other parts of Italy. It is therefore advisable to discuss the results from calibrated analysis along with the results of a traditional analysis.

Developments
The medical records algorithms were validated in another sample of GPs. PPV may be lower in the case of the GPs participating in this study. Moreover, absence of false negatives in the medical records was an assumption based on prevalence estimates. Both assumptions could be weakened: scenarios where sensitivity and specificity of medical records are imperfect could be designed and dealt with, using appropriate statistical methods [Gart1966, Valenstein1990].

Limitations
Our gold standard is *diagnosed* chronic disease. There is evidence that a substantial proportion of population with chronic conditions is not diagnosed, for instance for diabetes the figure could be up to 40% [Leong2013b].

CONCLUSION

For the three conditions that we investigated, case-finding algorithms on IAD records had excellent specificity and good PPV and NPV, but sensitivity lower than 75% and, in the case of IHD, lower than 50%. Longer look-backs of data are expected to improve these figures, but caution should be adopted in including too many years of look-back from drug utilization. Calibration can be used to assess the impact of imperfect case-finding algorithms on study results.

REFERENCES

[Altman1994] Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994;309(6947):102.

[Benchimol2011] Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. Journal of Clinical Epidemiology. 2011;64(8):821–9.

[Blais2014] Blais C, Dai S, Waters C, Robitaille C, Smith M, Svenson LW, et al. Assessing the Burden of Hospitalized and Community-Care Heart Failure in Canada. Canadian Journal of Cardiology. 2014;30(3):352-8.

[Brunelli2013] Brunelli SM, Gagne JJ, Huybrechts KF, Wang SV, Patrick AR, Rothman KJ, et al. Estimation Using All Available Covariate Information Versus a Fixed Look-back Window for Dichotomous Covariates. Pharmacoepidemiol Drug Saf. 2013;22(5).

[Buja2013] Buja A, Gini R, Visca M, Damiani G, Federico B, Francesconi P, et al. Prevalence of chronic diseases by immigrant status and disparities in chronic disease management in immigrants: a population-based cohort study, Valore Project. BMC Public Health. May 2013;13(1):504.

[Carstensen2008] Carstensen B, Kristensen JK, Ottosen P, Borch-Johnsen K. The Danish National Diabetes Register: trends in incidence, prevalence and mortality. Diabetologia 2008 Dec;51(12):2187-96.

[Chen2010] Chen G, Khan N, Walker R, Quan H. Validating ICD coding algorithms for diabetes mellitus from administrative data. Diabetes Research and Clinical Practice. Aug 2010;89(2):189–95.

[Chini2011] Chini F, Pezzotti P, Orzella L, Borgia P, Guasticchi G. Can we use the pharmacy data to estimate the prevalence of chronic conditions? a comparison of multiple data sources. BMC Public Health. Sep 2011;11:688.

[Corrao2011] Corrao G, Parodi A, Nicotra F, Zambon A, Merlino L, Cesana G, et al. Better compliance to antihypertensive medications reduces cardiovascular risk. J Hypertens. 2011;29(3):610–8.

[Cricelli2003] Cricelli C, Mazzaglia G, Samani F, Marchi M, Sabatini A, Nardi R, et al. Prevalence estimates for chronic diseases in Italy: exploring the differences between self-report and primary care databases. J Public Health Med. 2003;25(3):254–7.

[Flegal1986] Flegal KM, Brownie C, Haas JD. The effects of exposure misclassification on estimates of relative risk. Am J Epidemiol. 1986;123(4):736–51.

[Gart1966] Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. Am J Epidemiol. 1966;83(3):593–602.

[Gini2013] Gini R, Francesconi P, Mazzaglia G, Cricelli I, Pasqua A, Gallina P, et al. Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey. BMC Public Health. Jan 2013;13(1):15.

[Gini2014] Gini R, Schuemie MJ, Francesconi P, Lapi F, Cricelli I, Pasqua A, et al. Can Italian Healthcare Administrative Databases Be Used to Compare Regions with Respect to Compliance with Standards of Care for Chronic Diseases? PLoS ONE. 9 May 2014;9(5):e95419.

[Gini2015] Gini R, Schuemie MJ, Mazzaglia G, Lapi F, et al. Automatic identification of stages of type 2 diabetes, hypertension, ischaemic heart disease and heart failure from Italian General Practioners' electronic medical records: a validation study. Page 538 in: Abstracts of the 31st International Conference on Pharmacoepidemiology and Therapeutic Risk Management, August 22–26, 2015, Boston, Massachusetts, USA. Pharmacoepidemiol Drug Saf. Sep 2015;24:1–587.

[Giorda2012] Giorda C, Picariello R, Nada E, Tartaglino B, Marafetti L, Costa G, et al. The Impact of Adherence to Screening Guidelines and of Diabetes Clinics Referral on Morbidity and Mortality in Diabetes. PLoS ONE. Apr 2012;7(4):e33839.

[Gnavi2008] Gnavi R, Karaghiosoff L, Balzi D, Barchielli A, Canova C, Demaria M, et al. [Diabetes prevalence estimated using a standard algorithm based on electronic health data in various areas of Italy]. Epidemiol Prev. Jun 2008;32(3 Suppl):15-21.

[Gnavi2008b] Gnavi R, Karaghiosoff L, Costa G, Merletti F, Bruno G. Socio-economic differences in the prevalence of diabetes in Italy: The population-based Turin study. Nutrition, Metabolism and Cardiovascular Diseases. Dic 2008;18(10):678-82.

[Gnavi2011] Gnavi R, Canova C, Picariello R, Tessari R, Giorda C, Simonato L, et al. Mortality, incidence of cardiovascular diseases, and educational level among the diabetic and non-diabetic populations in two large Italian cities. Diabetes Res Clin Pract. May 2011;92(2):205-12.

[Gorina2011] Gorina Y, Kramarow EA. Identifying Chronic Conditions in Medicare Claims Data: Evaluating the Chronic Condition Data Warehouse Algorithm. Health Services Research. Oct 2011;46(5):1610–27.

[Green1983] Green MS. Use of predictive value to adjust relative risk estimates biased by misclassification of outcome status. Am J Epidemiol. 1983;117(1):98–105.

[Huber2013] Huber CA, Szucs TD, Rapold R, Reich O. Identifying patients with chronic conditions using pharmacy data in Switzerland: an updated mapping approach to the classification of medications. BMC Public Health. 2013;13(1):1030.

[Leong2013a] Leong A, Dasgupta K, Bernatsky S, Lacaille D, Avina-Zubieta A, Rahme E. Systematic Review and Meta-Analysis of Validation Studies on a Diabetes Case Definition from Health Administrative Records. PLoS ONE. 2013;8(10):e75256.

[Leong2013b] Leong A, Dasgupta K, Chiasson J-L, Rahme E. Estimating the Population Prevalence of Diagnosed and Undiagnosed Diabetes. Diabetes Care. 2013;

[Li2011] Li Q, Glynn RJ, Dreyer NA, Liu J, Mogun H, Setoguchi S. Validity of claims-based definitions of left ventricular systolic dysfunction in Medicare patients. Pharmacoepidemiology and Drug Safety. 2011;20(7):700–8.

[LoScalzo2009] Lo Scalzo A, Donatini A, Orzella L, Cicchetti A, Profi li S, Maresso A. Italy: Health system review. Health Systems in Transition, 2009; 11(6)1-216

[Magder1997] Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. Am J Epidemiol. 1997;146(2):195–203.

[Maio2005] Maio V, Yuen E, Rabinowitz C, Louis D, Jimbo M, Donatini A, Mall S, Taroni F. Using pharmacy data to identify those with chronic conditions in Emilia Romagna, Italy. J Health Serv Res Policy. 2005 Oct;10(4):232-8.

[Mazzaglia2009] Mazzaglia G, Ambrosioni E, Alacqua M, Filippi A, Sessa E, Immordino V, et al. Adherence to Antihypertensive Medications and Cardiovascular Morbidity Among Newly Diagnosed Hypertensive Patients. Circulation. 2009;120(16):1598-605.

[Morley2014] Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation. PLoS ONE. 2014;9(11):e110900.

[OECD2015] OECD. OECD Reviews of Health Care Quality. Italy 2014: Raising Standards. OECD Publishing; 2015.

[Poluzzi2005]Poluzzi E, Strahinja P, Vargiu A, Chiabrando G, Silvani MC, Motola D, et al. Initial treatment of hypertension and adherence to therapy in general practice in Italy. Eur J Clin Pharmacol. 2005;61(8):603–9.

[Quantin2013] Quantin C, Benzenine E, Velten M, Huet F, Farrington CP, Tubert-Bitter P. Self-controlled case series and misclassification bias induced by case selection from administrative hospital databases: application to febrile convulsions in pediatric vaccine pharmacoepidemiology. Am J Epidemiol. 2013;178(12):1731–9.

[Rogan1978] Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. Am J Epidemiol. 1978;107(1):71–6.

[Sakshaug2014] Sakshaug JW, Weir DR, Nicholas LH. Identifying diabetics in Medicare claims and survey data: implications for health services research. BMC Health Serv Res. 2014;14:150.

[Simonato2008] Simonato L, Baldi I, Balzi D, Barchielli A, Battistella G, Canova C, et al. [Objectives, tools and methods for an epidemiological use of electronic health archives in various areas of Italy]. Epidemiol Prev. Jun 2008;32(3 Suppl):5-14.

[Sultana2014] Sultana J, Italiano D, Spina E, Cricelli C, Lapi F, Pecchioli S, et al. Changes in the prescribing pattern of antidepressant drugs in elderly patients: an Italian, nationwide, population-based study. Eur J Clin Pharmacol. 2014;70(4):469-78.

[TheMatrix] TheMatrix. The tool for easy observational health data management from CSV files. http://thematrix.isti.cnr.it

[Valenstein1990] Valenstein PN. Evaluating diagnostic tests with imperfect standards. Am J Clin Pathol. 1990;93(2):252–8.

[Visca 2013] Visca M, Donatini A, Gini R, Federico B, Damiani G, Francesconi P, et al. Group versus single handed primary care: A performance evaluation of the care delivered to chronic patients by Italian GPs. Health Policy. Jun 2013

[Williams2012] Williams R. Using the margins command to estimate and interpret adjusted predictions and marginal effects. Stata Journal. 2012;12(2):308–31.

[WHO2016] WHO Collaborating Centre for Drug Statistics Methodology. ATC/DDD. International language for drug utilization research. http://www.whocc.no. Accessed June 2016.

ACRONYMS AND SHORT NAMES

ATC – Anatomical Therapeutic and Chemical classification system for drugs. It is an element of the ATC/DDD system of the World Health Organization.

DRUGS – list of drugs dispensed by community or hospital pharmacies free of charge or upon copayment. It is a table in the Italian Administrative Databases.

DDD – Defined Daily Dose of a drug. It is an element of the ATC/DDD system of the World Health Organization.

EXE – disease-specific exemptions from copayment to the healthcare system. It is a table in the Italian Administrative Databases.

HOSP – hospital discharge records. It is a table in the Italian Administrative Databases.

IAD – Italian Administrative Databases

IHD – ischaemic heart disease

LHU – Local Health Units

OUTPAT – outpatient healthcare dispensed by the healthcare system free of charge or upon copayment. It is a table in the Italian Administrative Databases.

PERSON – list of subjects who live in a defined geographical area. It is a table in the Italian Administrative Databases.

T2DM – type 2 diabetes mellitus

CHAPTER 4

Can Italian healthcare administrative databases be used to compare regions with respect to compliance with standards of care for chronic diseases?

Rosa Gini [1,2], Martijn J Schuemie [2], Paolo Francesconi [1], Francesco Lapi [3], Iacopo Cricelli [4], Alessandro Pasqua [4], Pietro Gallina [5], Daniele Donato [5], Salvatore Brugaletta [6], Andrea Donatini [7], Alessandro Marini [8], Claudio Cricelli [3], Gianfranco Damiani [9], Mariadonata Bellentani [10], Johan van der Lei [2], Miriam CJM Sturkenboom [2], Niek S Klazinga [11]

1. Agenzia regionale di sanità della Toscana, Florence, Italy
2. Department of Medical Informatics, Erasmus Medical Center, Rotterdam, The Netherlands
3. Società italiana di medicina generale, Florence, Italy
4. Genomedics, Florence, Italy
5. ULSS 16 Padova, Padua, Italy
6. ASP 7 Ragusa, Ragusa, Italy
7. Assessorato Politiche per la Salute, Bologna, Italy
8. Zona Territoriale Senigallia, Senigallia, Italy
9. Università Cattolica del Sacro Cuore, Rome, Italy
10. Agenzia Nazionale per il Servizi Sanitari Regionali, Rome, Italy
11. Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Published in
Gini R, et al. PLoS ONE 9(5): e95419. doi:10.1371/journal.pone.0095419
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0095419

ABSTRACT

Background

Italy has a population of 60 millions and a universal coverage single-payer healthcare system, which mandates collection of healthcare administrative data in a uniform fashion throughout the country. On the other hand, organization of the health system takes place at regional level, and local initiatives generate natural experiments. This is happening in particular in primary care, due to the need to face the growing burden of chronic diseases. Health services research can compare and evaluate local initiatives on the basis of the common healthcare administrative data.

However reliability of such data in this context needs to be assessed, especially when comparing different regions of the country. In this paper we investigated validity of healthcare administrative databases to compute indicators of compliance with standards of care for diabetes, ischaemic heart disease (IHD) and heart failure (HF).

Methods

We compared indicators estimated from healthcare administrative data collected by Local Health Authorities in five Italian regions with corresponding estimates from clinical data collected by General Practitioners (GPs). Four indicators of diagnostic follow-up (two for diabetes, one for IHD and one for HF) and four indicators of appropriate therapy (two each for IHD and HF) were considered.

Results

Agreement between the two data sources was very good, except for indicators of laboratory diagnostic follow-up in one region and for the indicator of bioimaging diagnostic follow-up in all regions, where measurement with administrative data underestimated quality.

Conclusions

According to evidence presented in this study, estimating compliance with standards of care for diabetes, ischaemic heart disease and heart failure from healthcare databases is likely to produce reliable results, even though completeness of data on diagnostic procedures should be assessed first.

Performing studies comparing regions using such indicators as outcomes is a promising development with potential to improve quality governance in the Italian healthcare system.

INTRODUCTION

Primary care is specifically suitable to face the growing chronic disease epidemic in a sustainable way [Starfield 2010, Fani Marvasti 2012] . Therefore it is the object of novel attention and of innovative policies [Barnes 2012] which specifically need health services research for timely effectiveness evaluation [Ettelt 2011, Klazinga 2011, Schafer 2011, Hansen 2011].

Many observational studies have been performed to evaluate the impact of innovative policies in primary care, for instance alternative rewarding policies for General Practitioners (GPs) in Ontario [Jaakkimainen 2011] or incentives for the introduction of Electronic Health Records in the United States [Cebul 2011, Reed 2012]. Such studies use administrative data to obtain evidence on the impact of policies in a inexpensive, timely and reproducible fashion [Solberg 2006]. Indicators measuring compliance with standards for management of chronic diseases were used as outcomes in those studies, similar to the clinical indicators of the Quality and Outcome Framework of the UK National Health System [NICE 2014], such as regular prescription of recommended therapies and regular diagnostic follow-up. However, concerns have been raised that such indicators estimated on the basis of administrative databases might not reflect the actual compliance of standards in the population bearing the disease, as methods to identify patients from administrative data, rather than clinical information, might lead to biased samples. Studies addressing this issue have obtained contradictory findings [Tang 2007, Green 2012].

As a result of those concerns, comparison of quality of primary care between regions or countries is generally performed by means of hospitalization rates for the so-called ambulatory care sensitive conditions [AHRQ 2007], which are readily obtained from administrative databases but do not require identification of cohorts of patients with a specific condition. However the relationships between quality of primary care and avoidable hospitalization is complex and population-based trends can be confounded by socioeconomic factors [Saxena 2006], by prevalence of morbidity or general hospitalization habits [Francesconi 2012].

In Italy, the VALORE Project was the first national-level study which evaluated a national policy in primary care by using administrative healthcare data for calculation of indicators of compliance [Visca 2013]. This paper presents the validation study on the reliability of administrative databases in estimating such indicators.

MATERIALS AND METHODS

Ethics Statement

No identifiable human data were used for this study. The dataset used in the study is not openly available. According to the Italian law on data confidentiality (decree 196/2003), permission to use individual-level data, albeit non-identifiable, must be granted by the institutions which bear the responsibility of the custody of the data. Permission to use data extracted from administrative databases for the VALORE project was granted to Agenzia regionale di Sanità della Toscana by ULSS 16 Padova (Veneto region), ASP 7 Ragusa (Sicily region), Assessorato Politiche per la Salute Emilia Romagna (Emilia Romagna region), Zona Territoriale Senigallia (Marche region), which are responsible for the custody of the data of the corresponding populations. Agenzia regionale di sanità della Toscana (Tuscany region) is enabled by a regional law (40/2005) to use Tuscan data for research purposes. Approval for use of encrypted and aggregated data from the HSD was also obtained from the Italian College of General Practitioners.

Setting

Italy has a tax-based, universal coverage national health system organised in three levels: national; regional (21 regions); and local (on average 10 Local Health Authorities per region). Healthcare is managed for every inhabitant by the Local Health Authority where she has her regular address [Lo Scalzo 2009]. Coordination of primary care within a Local Health Authority is performed at a smaller geographical level called Health District [Visca 2013]. Every Italian inhabitant is entitled to choose a GP, although parents might opt for a specialist paediatrician instead for their children, up to the age of 15. Therefore, each inhabitant from the age of 16 onward is specifically associated with a GP. GPs are the "gatekeepers" of the system, meaning that only upon GP prescription can specialist encounters be obtained free of charge. Dispensing of drugs or administration of diagnostic procedures can be obtained free of charge upon prescription of either a GP or a specialist physician employed by the healthcare system [Lo Scalzo 2009].

The five regions which contributed data to the VALORE validation study were: Veneto (A, Northern Italy), Emilia Romagna (B, Northern Italy), Tuscany (C, Central Italy), Marche (D, Central Italy) and Sicily (E, Southern Italy).

Study design

The VALORE project had selected several indicators to measure compliance with standards of care for diabetes, IHD and HF. In each region from the pool of regional GPs two convenience samples of groups of GPs were extracted and included in the validation study. In each regional pair, GPs of one sample had indicators computed from administrative databases, GPs of the other from their own clinical databases. Measurements of indicators were compared within and between regions.

The true values of an indicator across all the GPs in a region are an unobservable distribution. The rationale of this study design is based on the assumption that if measurements performed with two different methods in two different samples of GPs provide similar results, the likelihood that they both measure the true distribution is higher than the likelihood that they systematically make the same mistakes across different regions.

Data collection: sample of GPs with administrative- based measures

The national Italian government has mandated since the early Nineties collection of healthcare administrative data across the whole country. The healthcare activities which are mandated to be reported to the government have progressively expanded, from inpatient care [NSIS 1993] to drug dispensings and diagnostic tests [MINECO 2003]. Moreover an inhabitant registry is maintained by each Local Health Authority, where the GP chosen by the inhabitant is recorded, as well as other information, such as gender, birth date, date of entry in the territory of the Local Health Authority, date of exit from the territory [MINECO 2003]. However, outpatient diagnoses are not recorded in health administrative databases yet. Therefore cohorts of patients with chronic diseases must be selected by means of disease-specific longitudinal algorithms involving hospital discharges diagnoses, drug and/or other healthcare services utilization.

In each region, a convenience sample of Health Districts was chosen. All the GPs serving in those Health Districts were identified from the inhabitant registries of the corresponding Local Health Authorities and included in the sample. The healthcare administrative data of the whole population who chose a GP in this sample was loaded in the VALORE database. Patients aged 16-95 with diabetes, IHD and/or HF at the index date 1/1/2009 were detected by means of ad hoc algorithms based on past healthcare received. More details on this

process are described elsewhere [Gini 2013]. Indicators were computed during a one-year follow-up by linking the cohorts to the administrative databases of drug dispensings and diagnostic tests.
GPs were excluded from the samples if they had less than 300 persons registered or less than 4 patients with the disease, as indicators computed on small numbers were considered to be non robust.

Data collection: sample of GPs with clinical- based measures
The samples of GPs with clinical-based measures were drawn from the Health Search CSD Longitudinal Patient Database (HSD), a longitudinal observational database that is representative of the general Italian population. HSD was established in 1998 by the Italian College of General Practitioners and, at the time when the study was conducted, it contained data from from more than 800 GPs throughout Italy, covering a total population of around 1.2 million patients [Filippi 2005]. The GPs participating in HSD all use the same information software, in which they record demographic information, visits and referrals, diagnoses, drug and diagnostic tests prescriptions and clinical information of their patients. They are accepted as participants in HSD if their records are arguably complete, i.e. the prevalence of the principal diseases measured from their records is comparable with the expected prevalence of the general population. For this study, data from the 190 GPs practicing in the five regions of the VALORE project were used. The study population comprised patients aged 16-95 who had been registered with the GP for at least two years and were alive on 1st January 2009. Patients with diabetes, IHD and/or HF at the index date 1/1/2009 were detected by means of algorithms based on recorded diagnosis, which is described in detail elsewhere [Gini 2013]. Indicators were computed from the prescribed drugs or diagnostic tests.

Indicators
The indicators that were included in the study are shown in Table 1, and are classified as therapy indicators (for IHD and HF only), laboratory diagnostic tests, and bio-imaging diagnostic tests (HF only). All the indicators were based on clinical guidelines for the management of the disease that recommended regular therapy and yearly testing, respectively. The standard for a therapy recommendation was considered to be compliant with when at least two dispensings (in VALORE) or prescriptions (in HSD) were recorded in 2009,

at least 180 days the one from the other. The standard for a diagnostic recommendation (laboratory or bioimaging) was considered to be achieved when at least one test was performed (in VALORE) or prescribed (in HSD) during 2009.

Table 1. Indicators for the care of chronic diseases selected by the VALORE project and included in the validation study.

	Therapy (>1 dispensations per year, distance>180 days)	Laboratory diagnostic test (>0 per year)	Imaging diagnostic test (>0 per year)
Diabetes		Glycated emoglobin Creatinine	
IHD	ACE inhibitors Antithrombotics	Total colesterol	
IF	ACE inhibitors Beta-blockers		Ecocardiogram

Statistical analysis

In each sample the number of GPs, the number, age and gender distribution of patients aged 16+, and the average number of patients per GP were computed, both for the general population and for the population with each of the diseases. Differences in the variables within each regional pair of samples were tested either by a two-tail difference in means or a Chi-square test.

For each GP indicators were computed as percentage of patients who were compliant with the recommended standards of care. The distribution of the indicators of each regional pair were represented in a box-plot. To test whether each pair of measurements was drawn from the same distribution, the non-parametric Wilcoxon-Mann-Whitney two-sample statistic (also known as Wilcoxon rank-sum statistic) was performed in each region and for each indicator. In a sensitivity analysis, the test was repeated for achievements of standards in patients aged 45-74 years.

Data management and data analysis were performed with Stata 10.1.

RESULTS

Of the 1671 GPs serving in the Health Districts participating to the VALORE study, 1501 (89.8%) had enough registered patients and entered the study. Few GPs were discarded from the disease-specific studies because they had less than

four patients, the maximum was the 7% of GPs in region A in the HF study. All the 190 GPs in the HSD sample entered the study.

The description of the study population is shown in Table 2. Every HSD sample contained less GPs than the VALORE sample. The GPs in the HSD sample had a bigger registered population on average in all the five samples (range in HSD: 1238-1431, range in VALORE: 925-1223). The average number of patients per GP was higher in HSD GPs as well for diabetes (range in HSD 92.0-107.5, range in VALORE: 55.9-81.6) and IHD (range in HSD: 50.8-78.6, range in VALORE: 40.0-61.9), but for HF the numbers were similar in the two groups (range in HSD: 13.7-22.2, range in VALORE: 12.8-20.0). Age distribution was different within all pairs in all the populations, and the VALORE samples were older except in region B. Women were slightly more represented in the VALORE populations, except again in region B and in region E. This difference in gender did not show up in diabetic patients and was not consistent across regions in IHD and HF patients.

Figure 1 shows the box-plots of the pairs of distributions of the crude values of each indicator. A qualitative examination of the box-plots detected that distributions are very similar within the pairs. A notable exception are laboratory measurements in region E and bio-imaging test in all the regions, and VALORE showed lower values in all cases. The geographical trends, represented by orderings of the median values of the distributions, were similar between regions when measured in either data source, but less so in the case of the bio-imaging test.

Table 3 shows the results of the Wilcoxon-Mann-Whitney tests. Among therapy indicators the test found no differences in the distributions, with the exception of the samples in region C and, for HF only, region A, and the VALORE samples had higher values. The test confirmed that the distributions for all the laboratory diagnostic indicators of region E were different. Among diabetes the test detected slightly different distributions in three regions in either of the indicators, and in the IHD indicator region C and B had different distributions. The test confirmed that the only indicator of bio-imaging testing resulted in incoherent measurements in all but one region. Restricting the distributions to age-specific indicators (45-74) improved the comparability of the distributions of the therapy indicators of HF, and left unchanged the comparability of the other indicators.

Table 2. Description of the GPs in the five (A, B, C, D, E) pairs of samples (HS, measured by means of clinical data, and VALORE, measured by means of healthcare administrative data). Description of the general population they have in charge (general population), of diabetic patients (Diabetes), of patients with ischaemic heart disease (IHD) and of patients with heart failure (HF). N GP: number of GPs in the study. N populations: number of inhabitants registered with the GPs in the study. N patients: total number of patients with the chronic disease registered with GPs in the study. N registered per GP: average number of persons in charge to each GP, with test for difference in means within each pair. Female: percentage of women in the population, with test for difference in means within each pair. Age band: age distribution of the population, with chi square test within each pair.

			A			B			C			D			E		
			HSD	VALORE	P	HSD	VALORE	p	HSD	VALORE	P	HSD	VALORE	p	HSD	VALORE	p
General population	N GP		50	140		38	619		28	484		17	56		57	202	
	N population		67314	156516		54384	757163		34666	447577		23439	55992		75299	220843	
	N registered per GP		1346.3	1118.0	<0.001	1431.2	1223.2	<0.001	1238.1	924.7	<0.001	1378.8	999.9	<0.001	1321.0	1093.3	<0.001
	Female		52.7	54.5	<0.001	53.6	51.4	<0.001	52.5	54.2	<0.001	51.1	52.0	<0.05	52.8	50.9	<0.001
	Ageband	16-44	42.7	37.9	<0.001	37.4	42.3	<0.001	38.8	33.4	<0.001	43.5	35.6	<0.001	47.7	43.7	<0.001
		45-64	32.4	34.5		32.4	31.5		32.4	35.6		31.7	33.9		30.9	32.2	
		65-74	13.0	14.2		14.7	12.9		14.2	15.9		11.8	15.0		10.9	12.4	
		75-84	8.8	9.8		11.1	9.6		10.7	11.4		9.4	11.4		8.0	8.9	
		85-95	3.0	3.6		4.3	3.7		4.0	3.8		3.7	4.1		2.4	2.8	
Diabetes	N GP		50	137		38	619		28	484		17	56		57	202	
	N patients		4986	9122		4085	39138		2577	28872		1793	3132		7095	16490	
	N registered per GP		99.7	66.6	<0.001	107.5	63.2	<0.001	92.0	59.7	<0.001	105.5	55.9	<0.001	124.5	81.6	<0.001
	Female		45.2	46.7	0.089	46.0	47.1	0.193	48.3	49.3	0.313	46.4	47.5	0.454	51.4	50.8	0.388
	Ageband	16-44	5.2	6.4	<0.001	3.6	5.7	<0.001	4.0	5.2	<0.001	5.6	5.2	<0.05	7.0	4.3	<0.001
		45-64	32.0	25.5		29.2	27.3		31.8	26.6		31.7	26.8		37.0	30.5	
		65-74	32.3	30.4		31.4	28.7		30.4	30.1		28.7	29.7		27.9	29.9	
		75-84	22.8	27.2		26.9	27.1		25.9	28.5		24.7	27.6		22.7	26.9	
		85-95	7.7	10.6		8.9	11.2		7.9	9.6		9.4	10.8		5.5	8.4	

Table 2. Continued

			A			B			C			D			E		
			HSD	VALORE	P	HSD	VALORE	p	HSD	VALORE	P	HSD	VALORE	p	HSD	VALORE	p
IHD	N GP		50	137		38	619		28	484		17	56		57	201	
	N patients		2538	6041		2985	38324		1584	19366		1070	2622		3581	9782	
	N registered per GP		50.8	44.1	<0.05	78.6	61.9	<0.001	56.6	40.0	<0.001	62.9	46.8	<0.001	62.8	48.7	<0.001
	Female		38.1	41.9	<0.05	40.4	45.2	<0.001	39.4	42.0	<0.05	40.7	39.9	0.669	39.6	40.0	0.657
	Ageband	16-44	2.2	1.1	<0.001	1.1	1.1	<0.001	1.1	0.9	<0.001	1.3	1.1	<0.001	3.2	1.8	<0.001
		45-64	22.5	15.4		20.3	15.3		18.9	16.0		23.5	15.5		28.3	19.9	
		65-74	31.0	24.1		29.7	23.0		28.6	25.6		28.8	23.8		30.0	26.8	
		75-84	31.8	35.9		34.1	36.3		35.3	36.9		32.2	36.2		29.0	34.9	
		85-95	12.5	23.5		14.8	24.4		16.1	20.7		14.2	23.4		9.5	16.5	
HF	N GP		47	130		38	605		26	479		17	54		53	197	
	N patients		664	2110		842	12107		395	6577		296	692		726	2794	
	N registered per GP		14.1	16.2	0.080	22.2	20.0	0.150	15.2	13.7	0.229	17.4	12.8	<0.05	13.7	14.2	0.642
	Female		45.8	55.4	<0.001	56.3	51.8	<0.05	50.6	45.7	0.058	50.7	50.7	0.989	53.7	48.0	<0.05
	Ageband	16-44	0.5	0.7	<0.001	0.8	1.1	0.504	1.3	1.1	<0.001	1.4	0.3	0.253	1.7	2.0	0.671
		45-64	13.0	7.4		10.2	9.0		8.4	13.5		10.1	8.8		13.1	12.6	
		65-74	21.4	15.5		18.5	17.4		17.7	22.2		18.2	16.3		20.5	22.3	
		75-84	37.5	37.8		38.5	38.5		40.0	37.7		39.9	41.2		42.1	39.7	
		85-95	27.7	38.6		31.9	34.0		32.7	25.4		30.4	33.4		22.6	23.3	

Figure 1. Box-plots of the distribution of indicators of quality of care for diabetes (2 indicators), IHD (3 indicators) and HF (3 indicators) in 5 pairs of samples of GPs. Each pair contains the distribution obtained from the VALORE data (dark grey) and the distribution obtained from HSD data (light grey). For each indicator the pair of samples are ordered according to the median value in the VALORE sample.

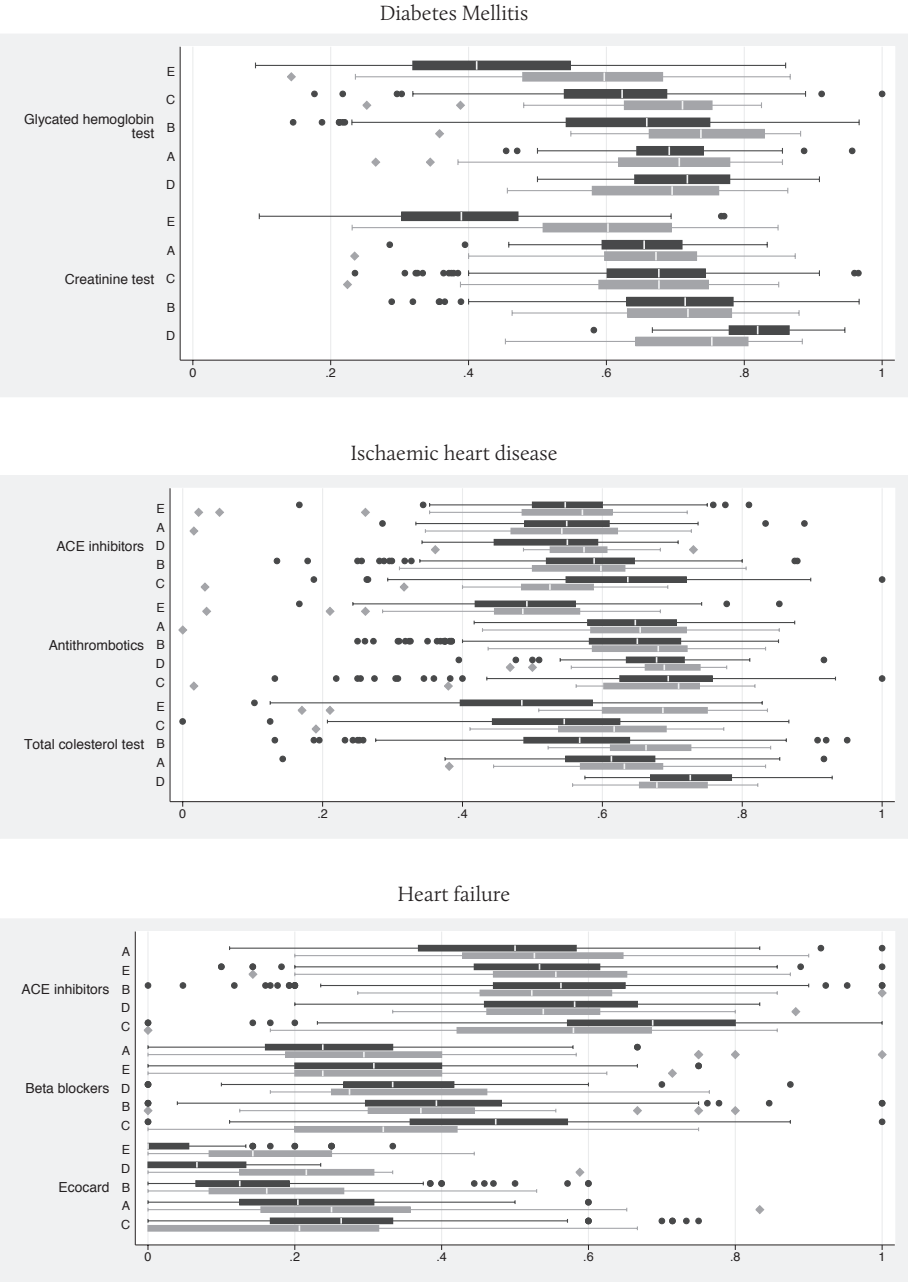


Table 3. P-values of the Wilcoxon-Mann-Whitney tests. P-values smaller than 0.05 are represented by a single star, P-values smaller than 0.001 are represented by a double star.

Disease	Indicators	Region	Pooled	45-74
Diabetes	Creatinine test	A	0.357	*
		B	0.587	0.701
		C	0.840	0.957
		D	*	*
		E	**	**
	Glycated emoglobin test	A	0.628	0.653
		B	**	*
		C	*	*
		D	0.441	0.441
		E	**	**
Ischaemic heart disease	ACE inhibitors	A	0.728	0.067
		B	0.695	0.671
		C	**	**
		D	0.116	0.065
		E	0.504	0.877
	Antithrombotics	A	0.508	0.174
		B	0.328	0.084
		C	0.651	0.440
		D	0.497	0.588
		E	0.754	0.095
	Total cholesterol test	A	0.225	0.962
		B	**	**
		C	*	**
		D	0.279	0.720
		E	**	**
Heart failure	ACE inhibitors	A	*	0.960
		B	0.454	0.107
		C	*	*
		D	0.701	0.961
		E	0.267	0.052
	Beta-blockers	A	*	0.670
		B	0.389	0.490
		C	**	0.091
		D	0.914	0.523
		E	0.293	0.614
	Ecocardiogram	A	0.134	0.245
		B	*	*
		C	0.059	0.944
		D	**	**
		E	**	**

DISCUSSION

Even though in Italy the data items to be collected in health administrative databases are mandated by the central government, and the resulting central databases are therefore formally homogeneous, data collection takes place locally. Italy is characterized by long-standing regional differences in general and in healthcare in particular [AIE 2011]. Therefore it is possible that inaccurate local data collection processes hamper data quality and completeness, and in particular quality of personal identifiers that allow for record-linkage. Moreover, outpatient diagnosis are not among the data items collected, therefore identification of cohorts of patients with chronic diseases must rely on algorithms linking inpatient diagnosis with drug and other healthcare services utilization. Inhomogeneous quality of personal identifiers and completeness of recordings might lead to inhomogeneous accuracy in defining cohorts of patients and in identifying healthcare services that they access. This in turn might result in non-comparable measures of compliance with standards of care for chronic diseases.

This study addressed this concern by comparing such measures with measures obtained from a different data source, in five Italian regions. The database which was chosen as a comparator collects clinical data from GPs, and is therefore complementary to the healthcare administrative data.

The results show that administrative databases provide reliable estimates on regional level. Indeed, the four therapy indicators had the same distribution within the pairs of regional samples in the large majority of cases. The same was observed for the three diagnostic indicators except in one region, where the distributions were systematically different. The only bio-imaging indicator had different distributions within pairs. Geographical trends between regions were consistent across the two data sources. This provides evidence that the two data sources both estimate the same population distribution, thus supporting the use of indicators computed on health administrative databases for comparisons between regions.

It was not possible to obtain measurements from the two data sources on the same samples of GPs. This was partly due to the fact that the identity of the GPs belonging to the database HSD is confidential. Moreover, data linkage at individual or even GP level between different data sources had legal implications in terms of privacy regulations and the procedures needed to obtain permissions

for such data collection [OCSE 2013] could not be managed in the context of the VALORE Project.

Therefore, observed differences in the distributions might be attributable to the composition of the following main effects: (a) due to non-random selection of the two samples, the GPs in the two samples were qualitatively different with actually different performance; (b) due to the different selection process that was conducted in the two type of data sources, the cohorts of patients of the two samples were qualitatively different subpopulations of the actual patients, which actually received different care; (c) difference in measurement, and HSD was likely biased (d) difference in measurement, and VALORE was likely biased. In the following paragraphs we provide plausible explanations to disentangle the effect (d), which is the object of this study, from (a), (b) and (c). It is a limitation of this study (see Limitations subsection) that some of the hypotheses we generated could not be tested. For cause (b), the main reference is the study by Gini et al, which found evidence that diabetic patients without therapy are less prevalent in the VALORE sample, and patients with heart failure are younger in the GP sample [Gini 2013].

For therapy indicators some differences are observed for HF in regions A and C. This is most likely due to reason (c), that is, patients included in the cohorts of HSD samples are different than patients included in the cohorts of the VALORE samples: indeed, age distribution of patients is different within the pairs, with the older cohort in VALORE being more likely to be assisted at home or in residential facilities, where GPs are likely to not record their activity completely [Filippi 2005, Gini 2013]. To test this assumption, analysis was restricted to patients aged 45-74, and indeed differences disappeared in region C in one indicator and in region A in both.

For laboratory testing indicators, region E seem to underestimate consistently the actual values of the indicators, across the three diseases. This could amount to incomplete collection of administrative data from laboratories, or to higher use of out-of pocket services: indeed, the most recent National Health Survey found that in region E (Sicily) attitude to use diagnostic services that are non reimbursed by the Health System is higher than in the other regions participating to our study [Rosano 2011]. In the other regions differences do not show a consistent pattern, except perhaps in region C, where however (a) rather than (d) could be the cause, that is, GPs in the HSD sample and GPs in the VALORE sample in region C actually have different quality of care. Indeed,

in region C therapy indicators differ slightly between samples as well.
The bio-imaging indicator is probably underestimated by healthcare administrative databases: this might be due to out-of-pocket payment of this analysis, or to the fact that bio-imaging occurring during hospital admissions was not recorded by VALORE.
The overall similarity in measurements that was observed in this study generates in turn two observations. First, the standards of care in the sample of GPs participating to the HSD database seem to be representative of the distribution of the whole population of GPs. This was surprising, as GPs in HSD are selected because of completeness in their recordings, and good recording habits were expected to be associated with better standards of care. The second observation is that specialist physicians who assist chronic patients are likely to involve GPs in regular prescription of therapies and diagnostic tests: indeed, if GPs were unaware of such prescriptions in the share of patients who are visited by a specialist, their clinical data would detect lower standards.
Our study was performed in samples drawn from regions belonging to three macro-areas of the country. Only a study performed in all regions could rule out the possibility that major issues show up in other areas, however the evidence we observed points to the direction of greater confidence. On the other hand, we do *not* claim that our results support reliability of similar measurements for *any* chronic disease. Indeed, this is determined by how reliable the algorithm for identifying the case is, and it was shown that this depends specifically on the disease, as frequency of hospital use, specificity of drug indication and pattern of healthcare may vary [Gini 2013].
In summary, the evidence we provide is promising enough to support comparison of regions with respect to indicators of compliance with standards of care for diabetes, IHD, and HF. Moreover, it supports the reliability of empirical studies, as the VALORE study [Visca 2013], using such indicators to evaluate the impact of organizational innovation in primary care.

Limitations
In this study indicators were computed on a population level for a convenience samples of GPs instead of directly being compared on a patient level for the same GPs. Similarity between samples could be due to random combination of contrasting effects rather than being attributable to the factors that we discussed. Although this is unlikely to have happened consistently in five

regions, an individual-level validation study only could address this concern. Italy, like several other countries, has a national legislation that permits exemption to the requirement for patient consent for projects in the public's interest [OECD 2013], but this pathway was too complex to be faced in the context of the VALORE project.

It was not possible to test some of the hypotheses we generated to explain observed differences. A study involving more regions and different points in time could provide counterfactuals to test our hypotheses.

CONCLUSION

According to the evidence presented in this study, estimating compliance with standards of care for diabetes, ischaemic heart disease and heart failure from healthcare databases is likely to produce reliable results, even though completeness of data on diagnostic procedures should be assessed first. Performing studies comparing regions using such indicators as outcomes is a promising development with the potential to improve quality governance in the Italian healthcare system.

REFERENCES

[AHRQ 2007] Agency for Healthcare Research and Quality. AHRQ quality indicators —guide to prevention quality indicators: hospital admission for ambulatory care sensitive conditions. Rockville, MD: AHRQ, 2007. (AHRQ Pub; no. 02-R0203

[AIE 2011] Costa G, Ricciardi W, Paci E (eds.) United Italy, 150 years later: has equity in health and healthcare improved? Epidemiol Prev 2011; 35 (5-6) suppl. 2: 1-136

[Barnes 2012] Barnes KA, Kroening-Roche JC, Comfort BW. The Developing Vision of Primary Care. New England Journal of Medicine. 2012;367(10):891–3.

[Cebul 2011] Cebul RD, Love TE, Jain AK, Hebert CJ. Electronic health records and quality of diabetes care. N. Engl. J. Med. 2011;365(9):825–33.

[Ettelt 2011] Ettelt S, Mays N. Health services research in Europe and its use for informing policy. Journal of Health Services Research & Policy. 2011;16(Supplement 2):48–60.

[Fani Marvasti 2012] Fani Marvasti F, Stafford RS. From Sick Care to Health Care — Reengineering Prevention into the U.S. System. New England Journal of Medicine. 2012;367(10):889–91.

[Filippi 2005] Filippi A, Vanuzzo D, Bignamini AA, Sessa E, Brignoli O, Mazzaglia G. Computerized general practice databases provide quick and cost-effective information on the prevalence of angina pectoris. Ital Heart J 2005; 6 (1): 49-51.

[Francesconi 2012] Francesconi P, Gini R, Maciocco G, Damiani G. [Primary care and chronic diseases: geographical differences in avoidable hospitalization]. Epidemiol Prev. 2011;35(5-6 Suppl 2):128–9.

[Gini 2013] Gini R, Francesconi P, Mazzaglia G et al. Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey. BMC Public Health. 2013;13(1):15.

[Green 2012] Green ME, Hogg W, Savage C, Johnston S, Russell G, Jaakkimainen RL, et al. Assessing methods for measurement of clinical outcomes and quality of care in primary care practices. BMC Health Services Research. 2012;12(1):214.

[Hansen 2011] Hansen J, Schafer W, Black N, Groenewegen P. European priorities for research on health care organizations and service delivery. Journal of Health Services Research & Policy. 2011;16(Supplement 2):16–26.

[Jaakkimainen 2011] Jaakkimainen RL, Barnsley J, Klein-Geltink J, Kopp A, Glazier RH. Did changing primary care delivery models change performance? A population based study using health administrative data. BMC Fam Pract. 2011;12:4

[Klazinga 2011] Klazinga N, Fischer C, ten Asbroek A. Health services research related to performance indicators and benchmarking in Europe. Journal of Health Services Research & Policy. 2011;16(Supplement 2):38–47.

[Lo Scalzo 2009] Lo Scalzo A, Donatini A, Orzella L, Cicchetti A, Profili S, Maresso A: In Italy: Health system review, Academic Press 2009:1-216. [Health Systems in Transition, 11(6)].

[NICE 2014] National Institute for Health and Care Excellence. NICE Menu of Indicators. http://www.nice.org.uk/aboutnice/qof/indicators.jsp. Accessed January 2014

[NSIS 1993] http://www.trovanorme.salute.gov.it/dettaglioAtto;jsessionid=TTE3ik3IdFrAc8xmb-Ujbw__?id=12531. Accessed October 2013

[MINECO 2003] http://www.lexitalia.it/leggi/dl_2003-269.htm. Accessed October 2013

[OECD 2013] OECD. Strengthening Health Information Infrastructure For Health Care Quality Governance: Good Practices, New Opportunities and Data Privacy Protection Challenges: Key Findings www.oecd.org/health/HealthPolicyBrief_OECD-Report-on-Health-Information-Infrastructure.pdf (accessed April 2013)

[Reed 2012] Reed M, Huang J, Graetz I, Brand R, Hsu J, Fireman B, et al. Outpatient Electronic Health Records and the Clinical Care and Outcomes of Patients With Diabetes Mellitus. Ann Intern Med. 2012;157(7):482–9.

[Rosano 2011] Rosano A, Giordani B. Out-of-pocket payment for specialistic care: copayment and critical issues] [http://www.asplazio.it/asp_online/att_territoriale/sias_new/pres_sias_2011/Rosano.pdf. Accessed January 2014

[Saxena 2006] Saxena S, George J, Barber J, Fitzpatrick J, Majeed A. Association of population and practice factors with potentially avoidable admission rates for chronic diseases in London: cross sectional analysis. JRSM. 2 gennaio 2006;99(2):81–9.

[Schafer 2011] Schäfer W, Groenewegen PP, Hansen J, Black N. Priorities for health services research in primary care. Qual Prim Care. 2011;19(2):77–83.

[Solberg 2006] Solberg LI, Engebretson KI, Sperl-Hillen JM, O'Connor PJ, Hroscikoski MC, Crain AL. Ambulatory Care Quality Measures for the 6 Aims From Administrative Data. American Journal of Medical Quality. 2006;21(5):310 –316.

[Starfield 2010] Starfield B. Reinventing primary care: lessons from Canada for the United States. Health Aff (Millwood). 2010;29(5):1030–6.

[Tang 2007] Tang PC, Ralston M, Arrigotti MF, Qureshi L, Graham J. Comparison of Methodologies for Calculating Quality Measures Based on Administrative Data versus Clinical Data from an Electronic Health Record System: Implications for Performance Measures. J Am Med Inform Assoc. 2007;14(1):10–5.

[Visca 2013] Visca M, Donatini A, Gini R, Federico B, Damiani G, Francesconi P, et al. Group versus single handed primary care: a performance evaluation of the care delivered to chronic patients by Italian GPs. Health Policy. Nov 2013;113(1-2):188–98.

CHAPTER 5

Monitoring compliance with standards of care for chronic disease using healthcare administrative databases in Italy: strengths and limitations

Rosa Gini [1,2], Martijn Schuemie [3,4], Alessandro Pasqua [5], Emanuele Carlini [6], Francesco Profili [1], Iacopo Cricelli [7], Patrizio Dazzi [6], Valentina Barletta [1], Paolo Francesconi [1], Francesco Lapi [5], Andrea Donatini [8], Giulia Dal Co [9], Modesta Visca [9], Mariadonata Bellentani [9], Miriam Sturkenboom [2], Niek Klazinga [10]

1. Agenzia regionale di sanità della Toscana, Osservatorio di epidemiologia; 50141 Florence, Italy
2. Department of Medical Informatics, Erasmus Medical Center; 3015 GJ Rotterdam, The Netherlands
3. Janssen Research & Development, Epidemiology; Titusville, New Jersey, United States
4. Observational Health Data Sciences and Informatics (OHDSI); New York, New York, United States
5. Società italiana di medicina generale, Florence, Italy
6. Consiglio Nazionale delle Ricerche, Istituto di Scienza e Tecnologie dell'Informazione, Pisa, Italy
7. Genomedics, Florence, Italy
8. Assessorato Politiche per la Salute, Bologna, Italy
9. Agenzia Nazionale per i Servizi Sanitari Regionali; 00100 Rome, Italy
10. Academic Medical Center, University of Amsterdam; 1100 DD Amsterdam, The Netherlands

Submitted for publication

ABSTRACT

Background

A recent comprehensive report on healthcare quality in Italy published by the Organization of Economic Co-operation and Development (OECD) recommended that regular monitoring of quality of primary care by means of compliance with standards of care for chronic diseases is performed. In a previous ecological study we explored whether compliance with standards of care could be reliably estimated on regional level using administrative databases. However, many questions remained.

Methods

We compared estimates of compliance with diagnostic and therapeutical standards of care for type 2 diabetes (T2DM), hypertension and ischaemic heart disease (IHD) from administrative data (IAD) with estimates from medical records (MR) for the same persons registered with 24 GP's. Data were linked at an individual level. For all the diseases MR was considered a gold standard for the denominator. MR and IAD were considered to contribute equally to the numerator of all the measures of compliance.

Results

32,688 persons entered the study, 12,673 having at least one of the three diseases according to at least one of the two data sources. The number of false negatives in IAD was relevantly high for all three conditions. IAD had imperfect sensitivity in detecting compliance to standards, especially in the case of diagnostic standards. Compliance with diagnostic standards was underestimated by IAD. The estimates of compliance with therapeutical standards were similar between IAD and MR. This finding can be partly explained through the coincidental combination of the high rates of false negatives and the low sensitivity on standards for IAD.

Conclusion

Using Italian administrative databases for purposes of quality monitoring is possible, but limitations should be noted. IAD seems a good source to monitor the quality of care, especially with respect to compliance with therapeutical standards. However, estimates are still flawed through the relatively high

numbers of false negatives and underestimation of compliance with standards. IAD can help signalling critical or excellent clusters at local or central level informing discussions on performance between GP's and local health unit managers and discussions between national and regional policy makers. An audit based on medical records is still the preferred method to assess a more comprehensive compliance with standards on the different levels. A close collaboration among GP's, managers and policy makers is needed to ensure further refinement of data collection, measurement of quality and accurate and actionable interpretation of results.

INTRODUCTION

A recent comprehensive report on healthcare quality in Italy published by the Organization of Economic Co-operation and Development (OECD) recommended that regular monitoring of quality of primary care by means of compliance with standards of care for chronic diseases is performed Italian National Healthcare System (NHS). Indeed, strengthening the national quality governance model on this sector of health care is a strategic objective in an ageing population, with an expected growing burden of chronic conditions. In the report, smarter payment systems for general practitioners that reward quality are advocated for, with specific reference to compliance with standards of care for chronic conditions [OECD2015].

However, measuring compliance with standards of care for chronic diseases is a challenging task for the Italian NHS [Gini2014]. Italian administrative databases (IAD) are available to the NHS uniformly from the whole country, and are the natural candidate data source. However their use poses a double problem: accuracy in the estimate of the denominator, and accuracy in the estimate of the numerator.

As for the denominator, the data items collected in IAD do not allow direct identification of patients with chronic conditions. Indeed, diagnoses performed in an outpatient setting are not collected in IAD, and this is generally the case when a chronic disease is diagnosed [Gini2013]. In Italy every adult patient is entitled to choose a general practitioner (GP), and specialist care can be requested to the NHS by patients only upon referral by their GPs. GP's soon become aware if a chronic disease is diagnosed in their patients. Primary care medical records (MR) rather than IAD may be the right source of information to define denominator.

On the other hand, for the numerator, compliance with standards of care may go undetected *both* by IAD and by MR. Over-the-counter purchase of prescribed drugs is not recorded in IAD, and drug prescriptions issued by specialists are not recorded by GPs. Diagnostic tests ordered by GPs or specialists are only recorded by IAD if they are performed in facilities belonging to, or contracted by, the NHS. This may fail to happen when access to such facilities is perceived as slow or cumbersome by patients and tests are performed outside the NHS system. Diagnostic exams are recorded in the MR either if GPs are the prescribers or if patients themselves provide the result to their GPs, since

there is no automatic transmission of test results in place in Italy. Ordering of diagnostic tests may more often be done by a specialist for more severe patients, or when the local organization of the healthcare system fails to encourage patients to access primary rather than secondary care. Hence sensitivity of MR in detecting diagnostic tests depends both on patient-level and on geographic-level characteristics [Gini2014].

For the reasons provided above, it was unclear whether compliance with standards of care in a population of patients with chronic diseases could be reliably estimated using IAD and, as a consequence, could be used to inform discussions on quality improvement and accountability on a local and central level. It was however evident that comparison with primary care MR had a chance to provide more knowledge on these questions.

In previous studies, case-finding strategies in IAD have been developed and validated [Gini2013, Gini2015b], and compliance with standards of care measured on this denominator has been compared with indicators obtained from a database of MR [Gini2014]. The results were encouraging, because estimates were very similar across the two data sources. However, the comparison was an ecological study, and many questions remained. It was not known whether the sample of patients detected by IAD was representative of the true set of patients with the disease, or rather false positives and false negatives in the denominator had different values of compliance. This could have combined with incompleteness of IAD in the numerator, to provide a falsely reassuring similarity between estimates.

The MATRICE Project, started by Italian National Agency for Regional Health Systems in 2011, aimed to assess in a more comprehensive way the validity of IAD as a data source to monitor quality of health care for chronic diseases. MATRICE obtained from the National Authority for Personal Data Protection permission to link IAD and MR of a large sample of patients. In this study we could therefore compare compliance with standards of care for type 2 diabetes mellitus (T2DM), hypertension and ischaemic heart disease (IHD) using both data sources, at the individual level.

METHODS

Study design

On each subject we defined denominators and numerators of each standard of care, both using IAD and using MR. Based on a previous study, we considered MR to be a gold standard for the denominator [Gini2015a, Gini2015b]. Therefore, for each disease, we considered as true positives the persons who were in the denominator for both sources, as false positives the persons who were in the denominator for IAD, but not for MR, and as false negatives the persons who were in the denominator for MR, but not for IAD. Based on the arguments in the Introduction, we considered that neither IAD nor MR had complete information on the numerator, so we assessed concordance among the two variables, in the whole study population. To assess concordance among measures of indicators, and representativeness of the IAD denominators, we compared the indicators (ratio of numerator to denominator) in several subpopulations (denominators according to IAD, denominator according to MR, true positives, false positives, false negatives) using as numerator in turn IAD, MR and either of the two sources.

See the subsection "Study variables" below for more details on how the variables were defined, and the subsection "Data analysis" for more details on the statistical analysis.

Setting

From the point of view of organization of health care, Italy is divided into 21 regions, and each region is divided in geographic subareas (on average 10 per region). Health care for the population in each area is managed by organizations called Local Health Units (LHU). LHUs collect administrative data on the health care they provide to their inhabitants which together form the basis of the IAD. The National College for General Practitioners (SIMG) is the national scientific society of General Practitioners in Italy, which provides training to the GPs to improve the quality of recording in their medical records.

A sample of 25 GPs belonging to five regions was recruited in this study. Three regions were in the North (Lombardy, Veneto, Emilia-Romagna), one in the Center (Tuscany) and one in the South (Puglia). All the GP's of the same region belonged to the same LHU.

Standards of care

A panel of experts in organization of primary care services, epidemiologists and clinicians selected clinical guidelines for T2DM, hypertension and IHD which were expected to be easy to monitor on IAD. The result is depicted in Table 1: six indicators for annual diagnostic follow-up and treatment with four drug classes were chosen, each applying to one or more of the three conditions, totalling 18 indicators. Each recommendation is labelled with the name of the scientific society who published it, and with its grade and level [GRADE2004].

Data collection

A script was developed by SIMG to automatically query the medical records of the 25 GPs. The script first identified all subjects in charge to the GP at 1st January 2012. Then it computed variables estimating compliance with the standards of care during 2012 for each subject. Finally it applied validated algorithms to detect whether subjects had one or more of the diseases under study [Gini2015a].

All the IAD data available to the healthcare system on the same population was extracted from the LHUs, using TheMatrix, an open source software tool [TheMatrix].

Personal identifiers were pseudonymized at extraction, using the same encryption key, and all the data was automatically transmitted to the National Research Council (CNR), which had been granted permission to store and process this data. Investigators from Agenzia regionale di sanità della Toscana (ARS) developed a script to compute the study variables from IAD data, and CNR ran it on the IAD data. Finally, CNR linked the analytical dataset and medical records at individual level and transmitted the resulting dataset to ARS for statistical analysis.

One of the GPs from Lombardy was on leave in 2012 and was therefore discarded from the study after data collection.

Study variables

Case-finding algorithms to compute denominators for T2DM, hypertension and IHD from MR were selected based on a previous validation study. This study proved that the case-finding algorithms of the three diseases all had almost perfect positive predictive value [Gini2015a]. Since population prevalence estimated with those algorithms are very high, sensitivity must be very high

as well [Gini2013]. For this reason, in this study we used the denominator from MR as a perfect identification of the true denominator. Case-finding algorithms for denominators from IAD used a combination of diagnosis from hospital discharge records, disease-specific exemptions from copayment, and utilization of treatments. These algorithms are described in detail in Chapter 3 of this thesis. Sensitivity and positive predictive values of those algorithms were estimated in a separate study [Gini2015b].

Numerators were defined similarly across the two data sources. The subject was considered to be compliant with a treatment if at least two records of the treatment with different dates were found in 2012, and compliant with a recommended diagnostic test if at least one prescription for that test was found in 2012, except in the case of glycated hemoglobin where two records were requested.

As discussed in the Introduction, both IAD and MR have imperfect sensitivity to define the numerators. For each standard of care we analysed three different variables: numerator as measured by IAD, numerator as measured by MR, numerator as measured by either source (EITHER).

Measures of compliance

For each person in the study population we had variables estimating denominators and variables estimating numerators computed both from IAD and from MR. We were therefore able to compare three ways of estimating the ratio between numerator and denominator; based on IAD only, based on MR only, and based on either IAD or MR. When based on IAD only: both denominator and numerator are estimated from IAD. This is mainly the perspective from the national and regional NHS policy maker, who have only IAD data available. When based on MR only, the denominator and numerator both come from MR. This is usually the perspective of the GP when evaluating his/her own practice. A third perspective takes the whole set of services used by the population into account. This perspective, the true value of compliance with the standard of care in the population with the disease, is often lacking. With our data we could estimate this measure by using denominator from MR and numerator from either MR or IAD: MR is the best possible denominator, because it is a gold standard, and "either MR or IAD" is the best possible numerator, because the two data sources compensate each other's incompleteness. We refer to this as the "best possible estimate".

Quality governance scenarios

We considered two scenarios of quality governance where the results of our comparison can be useful, as shown in Table 1: a *local* scenario, when local or regional decision makers discuss quality of care with GPs with a focus on quality improvement, and a *central* scenario, when regional or national decision makers discuss quality of care, respectively, with local or regional decision makers with a focus on quality monitoring. To support the local governance we compared the point of view of the healthcare system with the point of view of GP's, on clusters of patients assisted by the same GP. To support the central governance we compared the point of view of the healthcare system with the estimated population-based value on clusters of patients assisted by the same LHU.

Table 1. Standards of care, with levels and grades of recommendation. SID: Italian Diabetes Society. ESC/EASD: European Society of Cardiology and European Association for the Study of Diabetes. ACC/AHA: American Cardiology Association and American Heart Association. A symbol * means that the recommendation only applies when the condition is at a high level of severity. Diagnostic tests are recommended once per year, except HbA1c for T2DM which is recommended twice a year.

Guideline type	Recommendation in the guideline	T2DM	Hypertension	IHD
Therapeutic	Statins	level I, grade A **(SID)**		Level IIa, grade B **(ACC/AHA)**
	Beta-blockers			Level I, grade A **(ACC/AHA)** *
	ACE inhibitors			Level I, grade A **(ACC/AHA) ***
	Antithrombotics			Level I, grade A **(ACC/AHA)**
Diagnostic	Microalbuminuria test	level VI, grade B **(SID)**	level I, grade B **(ESH/ESC)**	
	HbA1c tests	level VI, grade B **(SID)**	level I, grade B **(ESH/ESC) $**	Level I, grade A **(ESC/EASD)**
	Lipid profile	level III, grade B **(SID)**	**(ESH/ESC)**	Level III, grade B **(ACC/AHA)**
	Clearence/ creatinine test	level VI, grade B **(SID)**	level I, grade B **(ESH/ESC)**	
	ECG		level I, grade B **(ESH/ESC)**	X*
	Eye exam	level III, grade B **(SID)**	level IIa, grade C **(ESH/ESC)**	

Data analysis

For each disease we identified the denominator according to IAD, and computed false positives and false negatives using MR as a gold standard.

Since we didn't have a gold standard for numerators, we computed Cohen's kappa between the MR and IAD numerators. Concordance was categorized as "Poor" (<0.20), "Fair" (0.21-0.40), "Moderate" (0.41-0.60), "Good" (0.61-0.80), "Very good" (0.81-1.00) [Landis1977]. Moreover, we computed the percentage of those in the numerator in one source that overlaps with those in the numerator in the other source. We computed the increase in the numerator when adding EITHER to IAD.

For each standard and each cluster (GP or LHU), the three measures (ratios of numerator to denominator) were standardised per age and gender, using as a standard the age and gender distribution obtained from MR (data not shown).

We estimated average difference between the indicators computed by pairs of sources on each cluster and tested significance. Estimates were obtained by fitting logistic models on a dataset with a record per patient and source, with the source of information (IAD vs MR, or IAD versus EITHER) as a dependent variable. Variance was estimated by clustering the observations on the same subject. Models were adjusted per LHU, age band and gender, with interaction between source and cluster variable (GP or LHU).

For each disease, to assess whether patients in the IAD denominator (true positives + false positives) were representative of the true population with the disease (true positives + false negatives), we estimated the difference between compliance computed on true positives and, respectively, false positives and false negatives. In this analysis compliance was estimated with EITHER, and was adjusted per LHU, age and gender.

Ethics

Permission to perform record linkage between pseudonymized administrative data and medical records was granted by the Italian National Authority for the Privacy regulation. Specifically, permission was granted to CNR to store and process the data, and to ARS to obtain the linked individual-level analytical dataset, for statistical analysis.

RESULTS

Study population
Data on 32,688 subjects was collected. The average number of patients per GP was 1,362 (IQ range: 1,209-1,500).

Comparison of variables detecting diseases (denominators)
12,673 subjects had at least one of the three diseases according to at least one of the two data sources. According to IAD, 2,047 subjects had T2DM: only 107 (5%) were false positives, but additional 823 subjects were false negatives, according to MR (+40%). 8,392 subjects had hypertension according to IAD: 1,103 (13%) were false positives, and additional 3,573 subjects were false negatives, according to MR (+42%). 745 subjects had IHD according to IAD: 145 (19%) were false positives, and additional 776 subjects were false negatives (+ 104%).

Comparison of variables measuring compliance (numerators)
On the general population Cohen's kappa showed very good concordance (from 0.92 to 0.89) in the four indicators of compliance with therapies. Among diagnostic tests, concordance was very good (0.84) for microalbuminuria, good (from 0.76 to 0.66) for glycated hemoglobin, lipid profile and creatinine, moderate (0.44) for ECG and fair (0.27) for eye exams (0.27) (Table 3). Information provided by MR was almost complete (from 97% to 94%) for compliance with therapies, and was more complete than IAD in all the other indicators except eye exam (20%) (Table 3). Adding EITHER to IAD increased the numerator by less than 15% in the case of therapies and of eye exam, from 24% to 32% in microalbuminuria, glycated hemoglobin, creatinine and lipid profile, and more than 50% for ECG.

Comparison of indicators of compliance
Scatter plots of the age-and-gender standardized indicators on the clusters of patients are represented in Figure 2.
IAD and MR had on average very similar estimates for therapeutic indicators, although for statins in both T2DM and IHD, and for betablockers in IHD, IAD had a significantly higher estimate (respectively +4.1, +4.5 and +5.4). The results were confirmed when comparing IAD with the "best estimate", and differences were reduced. In the case of diagnostic indicators, the picture was

Figure 1. Scatter plots comparing age-and- gender standardised measures of compliance with standards of care, in the two governance scenarios. In the Local governance scenario the 24 clusters of patients of the same GP are measured by IAD on the Y-axis and MR on the X-axis. In the Central governance scenario the 5 clusters of patients in the same LHU are measured by IAD on the Y-axis and best estimate (MR denominator and EITHER numerator) on the X-axis.

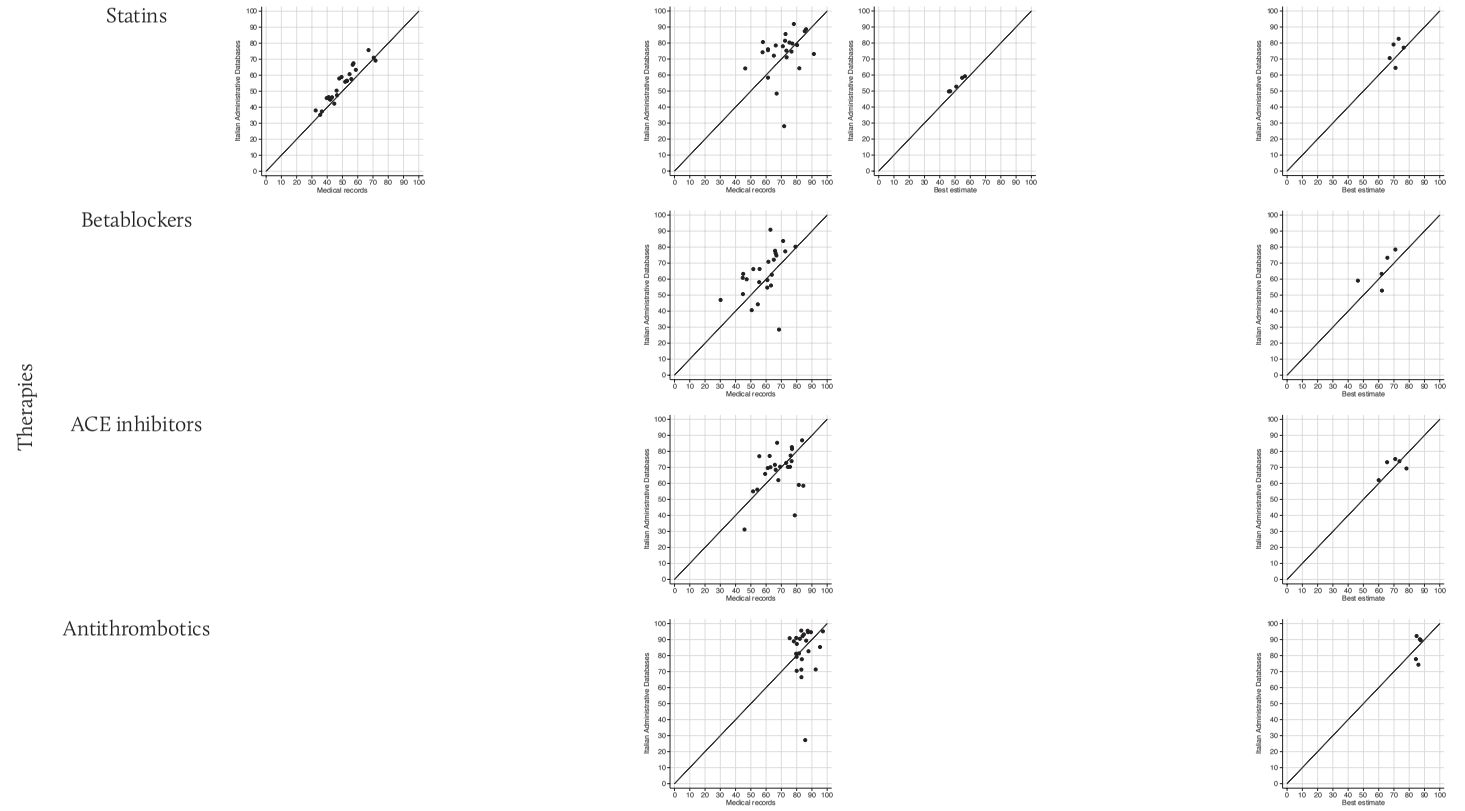


Figure 1. Continued
Local governance
Central governance
Type 2 Diabetes
Hypertension
Ischaemic heart disease
Type 2 Diabetes
Hypertension
Ischaemic heart disease
Diagnostic tests
Microalbuminuria
Glycated hemoglobin
Lipid profile
Creatinine
Italian Administrative Databases
Medical records
Best estimate

Figure 1. Continued

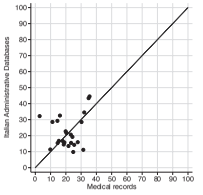
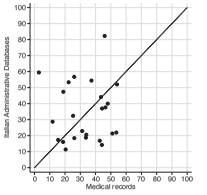
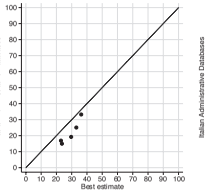
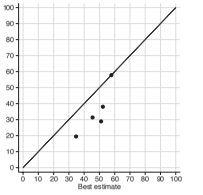
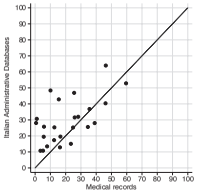
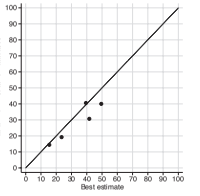
		Local governance			Central governance		
		Type 2 Diabetes	Hypertension	Ischaemic heart disease	Type 2 Diabetes	Hypertension	Ischaemic heart disease
Diagnostic tests	ECG						
	Eye exam						

Table 2. Scenarios of quality governance where the results from this study can be used. IAD: administrative databases, MR: primary care medical records, EITHER: either among IAD or MR

Quality governance scenarios	Local		Central	
Activity	Quality improvement		Quality monitoring	
Actors	Local decision-makers	GPs	Local (regional) decision-makers	Regional (national) decision-makers
Clusters	Patients assisted by the same GP		Patients assisted by the same LHU (region)	
Point of view	Healthcare system	GP	Healthcare system	Best estimate
Denominator	IAD	MR	IAD	MR
Numerator	IAD	MR	IAD	EITHER

more complex, with IAD showing higher values than MR and lower values than the "best estimate", often significantly. Average difference between IAD and MR was significant and higher than 5 percentage points for glycated hemoglobin and eye exam in T2DM. Average difference between IAD and "best estimate" was significant in all indicators except glycated hemoglobin, and in all but microalbuminuria and eye exam (Table 4).

Representativeness of subpopulations

Indicators in false negatives were much lower (from 15.8 to 40.1 percentage points difference) with respect to indicators in true positives in the case of T2DM, and substantially lower (from 6.7 to 24.6 percentage points difference) in the case of IHD (Table 3). They were lower in the case of hypertension, too, but less so (from 4.8 to 13.9 percentage points difference). Differences were higher for indicators of therapies. Differences between true positives and false positives were similar but slightly smaller (Table 3).

Table 3. Comparison of numerators measured by IAD, by MR or by either of the two data sources, on the whole population. Difference of indicators between true positives and false negatives (FN), and between true positives and false positives (FP). Difference was computed using EITHER for numerator, and adjusting per age, gender and LHU.

*Standards are listed in decreasing order of Cohen's kappa.

					Difference of indicators between true positives and false negatives (FN), and between true positives and false positives (FP)					
	Cohen's K	Percentage of those in the numerator in one source that overlaps with those in the numerator in the other source.		Percentage increase in the numerator when adding EITHER to IAD	T2DM		Hypertension		IHD	
Indicator *		Of those in IAD numerator	Of those in MR numerator		FN	FP	FN	FP	FN	FP
Statins	0.92	97.0%	89.4%	+11.5%	-19.3	-18.5			-24.6	-18.2
Betablockers	0.91	95.7%	88.8%	+12.1%					-21.9	-16.2
ACE inhibitors	0.90	96.6%	88.0%	+13.2%					-11.7	-6.9
Antithrombotics	0.89	94.5%	86.6%	+14.7%					-15.0	-12.8
Microalbuminuria	0.84	93.3%	78.1%	+26.2%	-38.6	-29.0	-13.3	-16.8		
Glycated hemoglobin	0.76	77.3%	76.6%	+23.7%	-40.1	-32.8	-4.8	-10.3	-6.7	-4.2
Lipid profile	0.73	91.0%	73.9%	+32.1%	-26.6	-13.5	-13.9	-16.3	-9.1	-2.2
Creatinine	0.66	84.2%	74.7%	+28.5%	-15.8	-13.9	-7.7	-13.9		
ECG	0.46	52.1%	50.8%	+50.5%			-5.9	-12.3	-6.5	-6.3
Eye exam	0.27	20.1%	62.5%	+12.1%	-20.1	-14.8				

Table 4. Average difference between the indicators computed in pairs of sources. On the left (local governance scenario): comparison between MR and IAD. On the right (central governance scenario): comparison between IAD and "best estimate". For each indicator the p-value of the significance of the difference is shown.

		Local governance scenario						Central governance scenario					
		T2DM		Hypertension		IHD		T2DM		Hypertension		IHD	
	Indicator	Δ	p	Δ	p	Δ	p	Δ	p	Δ	p	Δ	P
Therapeutic	Statins	4.1	<0.001			4.5	<0.05	3.3	<0.001			3.6	<0.05
	Betablockers					5.4	<0.001					3.3	<0.05
	ACE inhibitors					2.2	0.152					0.0	0.992
	Antithrombotics					1.3	0.345					-0.5	0.691
Diagnostic	Microalbuminuria	-0.1	0.931	-0.2	0.543			-3.4	<0.001	-2.3	<0.001		
	Glycated hemoglobin	8.9	<0.001	2.2	<0.001	3.0	<0.05	-0.1	0.929	-0.0	0.950	-1.1	0.366
	Lipid profile	-3.0	<0.001	-3.9	<0.001	-2.9	0.074	-6.3	<0.001	-7.2	<0.001	-6.9	<0.001
	Creatinine	0.8	0.402	-1.2	<0.05			-5.7	<0.001	-7.4	<0.001		
	ECG			0.1	0.891	1.2	0.539			-7.7	<0.001	-14.1	<0.001
	Eye exam	7.5	<0.001					-4.7	<0.001				

DISCUSSION

Numerators in the whole population were concordant between MR and IAD in the case of therapies, less so in the case of diagnostic testing, especially when more complex tests were considered (ECG and eye exam). Indicators of compliance with therapies showed low average difference between data sources, although still significant in some cases. Indicators of compliance with diagnostic monitoring were imbalanced: IAD estimated higher compliance with respect to MR, and lower compliance with respect to the best possible estimate. This was the result of a combination of different errors. Patients in the denominators for IAD were not representative of the true population of patients, especially in the case of T2DM and of therapeutic indicators. Small average differences between the estimates of IAD and the best estimates are partially coincidental and therefore run the risk of not being reproducible in all the regions and across time.

Interpretation of the findings: estimating compliance with recommended therapies and diagnostic tests

This individual-level study showed that the confounding effects anticipated in the limitations of the ecological study were indeed playing an important role in the estimate of indicators performed on IAD.

The effects of different misclassifications were balanced in the case of therapies, because concordance between MR and IAD was high, and MR was almost complete; therefore, the absence from the denominator of false negatives, who had lower compliance, compensated the small overall underestimation of the numerator. A small contribution was also provided by the comparatively small share of false positives, who had similar profile as false negatives.

As expected, numerators from IAD and MR were less concordant in the case of diagnostic tests, and numerators measured by IAD were lower. The combination of errors produced both balanced and imbalanced results. In the case of glycated hemoglobin test, in false negatives the indicator was less that 40 percentage points lower than in true positives and the overall agreement between the administrative and "best estimate" was due to underestimation of the numerator on the IAD denominator. However there was in important imbalance between IAD and MR estimate in T2DM patients. In recent, similar validation studies of estimates of measures of performance on diabetic patients

from administrative databases from the United States, similar mixed effects were observed [Hirsch2013, Sakshaug 2014].

Consequences on the use of Italian administrative data in a systematic quality monitoring and improvement strategy

In a quality monitoring strategy IAD seems a reliable tool for signalling purposes: when IAD detects either an excellent or a poor performance in a cluster of patients, according to our data it is very likely that the observation corresponds to a truly interesting fact, particularly in the case of compliance with therapeutic standards, and with yearly laboratory diagnostic tests.

However, we found that a combination of mutually balancing misclassifications is at the root of the similarity between IAD results and our best estimate of the true compliance in the patients with a diagnosed chronic disease, especially in the case of diagnostic recommendations. Specific caution should be taken in interpreting coverage of the twice-yearly glycated hemoglobin test in diabetic patients. Likewise, the measures of compliance with annual eye exam in diabetics, and annual ECG in hypertensive and IHD patients look fragile.

This has slightly different implications for a "local" quality improvement rather than a "central" quality monitoring scenario.

In a local scenario the main actors are, on the one hand, the local (or regional) decision-makers for the organization of healthcare for chronic diseases and, on the other, the GPs. The main objective is promoting appropriateness in healthcare for chronic diseases, that is, supporting the role of primary care as the main driver, in close collaboration with specialist care [OECD2015]. Thanks to IAD, decision makers have the possibility of producing estimates of compliance across a range of GPs. While this sort of comparison is in itself very informative, it is clear from our validation that it is not precise enough to provide a reliable ranking of the performance of GPs, nor to support quality-based payment systems, such as a pay-for-performance scheme. Rather it should be taken as the starting point for quality improvement initiatives, such as a more detailed audit of quality based on medical records. Clusters of patients with low compliance, as signalled by IAD, must be analysed in conjunction with context information, such as accessibility to local NHS facilities for diagnostic tests, and possible drive of local specialist healthcare providers towards replacing, rather than supplementing, primary care, sometimes implying out-of-pocket purchase of care. Both those elements can provide input to action for local decision-makers.

Clusters with high compliance, in turn, must be critically analysed: if patients with mild forms of chronic diseases are not appropriately followed-up, they will remain undetected by IAD, which will therefore measure higher compliance only on the more severe patients, thus providing a falsely reassuring picture. This is likely to be associated with clusters where IAD detects low prevalence. Aside from those extremes, quality governance at the local levels should focus on an integrated interpretation of IAD and MR data, which are both available to the actors.

In a central scenario the main actors are all decision-makers for the organization of healthcare system, at different levels: local vs regional, or regional vs national. The main objective is monitoring quality of healthcare and making comparisons between the different geographical entities to assure equity in quality of care amongst the whole Italian population. Integrated analysis of IAD and MR is not possible in this scenario, therefore context for interpretation of signals from IAD must be carefully built in collaboration with local decision-makers, who can provide crucial context information, in particular findings from local analysis of MR. Several resources are available to inform this assessment: SIMG produces a yearly report comparing compliance with standards of care across Italian regions estimated from MR of a sample of GPs [HS2015], and survey data are produced every five years by the National Institute of Statistics, estimating access to NHS specialist facilities [ISTAT2015].

Developments

Routine data-linkage between administrative data and key elements from primary care medical records, such as diagnosis of a chronic disease and compliance with standards of care, would critically improve the quality governance of primary care. Local initiatives have been initiated to this respect, such as the SOLE network in Emilia-Romagna [OECD2015].

Analytical calibration methods that include the results from this validation study, as well as aggregated measures produced by SIMG and the National Institute of Statistics, could be developed to improve estimates produced by IAD.

Implications for the use of the indicators in studies of impact

Indicators of compliance with standards of care can be used to evaluate the impact of innovative strategies [Chien2012, Visca2013]. Our results support

overall this use of the indicators, provided a difference-in-differences design is adopted, and the impact is measured across a short time span, so that it can be assumed that misclassification does not change differentially across exposed and non exposed to the intervention. If this is not possible, elements that may imbalance misclassification across exposed and non-exposed, or across time, need to be discussed in the limitations of the study.

Permission to perform record linkage was an extraordinary result
This study was made possible by an explicit permission of the Italian National Authority for the Privacy regulation, which allowed individual-level record linkage between IAD and MR on a large sample of patients. It is encouraging that such permission was granted, and routes for expedited permission should be created, especially for validation studies of administrative data. Indeed, this would allow rapid generation of evidence crucial for public health and health system governance in a transparent and legal manner.

Limitations
The numerator that we used as a "best estimate" may have overestimated the true compliance, as GP drug prescriptions may have not been filled in, and GP test orders may have not been performed in reality. The first effect is however likely to be small, as a second prescription is required for the patient to be compliant. Moreover, the concordance we observed between MR and IAD data was very high in numerators of therapeutic standards (Table 3).

CONCLUSION

Using Italian administrative databases for purposes of quality monitoring is possible, but limitations should be noted. IAD seems a good source to monitor the quality of care, especially with respect to compliance with therapeutical standards. However, estimates are still flawed through the relatively high numbers of false negatives and underestimation of compliance with standards. IAD can help signalling critical or excellent clusters at local or central level informing discussions on performance between GP's and local health unit managers and discussions between national and regional policy makers. An audit based on medical records is still the preferred method to assess a more

comprehensive compliance with standards on the different levels. A close collaboration among GP's, managers and policy makers is needed to ensure further refinement of data collection, measurement of quality and accurate and actionable interpretation of results.

REFERENCES

[Chien2012] Chien AT, Eastman D, Li Z, Rosenthal MB. Impact of a pay for performance program to improve diabetes care in the safety net. Preventive Medicine. 2012;55, Supplement:S80–5.

[Gini2013] Gini R, Francesconi P, Mazzaglia G et al. Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey. BMC Public Health. 2013;13(1):15.

[Gini2014] Gini R, Schuemie MJ, Francesconi P, Lapi F, Cricelli I, Pasqua A, et al. Can Italian Healthcare Administrative Databases Be Used to Compare Regions with Respect to Compliance with Standards of Care for Chronic Diseases? PLoS ONE. 2014;9(5):e95419.

[Gini 2015a] Gini R, Schuemie MJ, Mazzaglia G, Lapi F, et al. Automatic identification of stages of type 2 diabetes, hypertension, ischaemic heart disease and heart failure from Italian General Practioners' electronic medical records: a validation study. Page 538 in: Abstracts of the 31st International Conference on Pharmacoepidemiology and Therapeutic Risk Management, August 22–26, 2015, Boston, Massachusetts, USA. Pharmacoepidemiol Drug Saf. Sep 2015;24:1–587.

[Gini 2015b] Gini R, Schuemie MJ, Pasqua A, Dazzi P, et al. Identifying Chronic Conditions from Data Sources with Incomplete Diagnostic Information: The Case of Italian Administrative Databases. Page 538 in: Abstracts of the 31st International Conference on Pharmacoepidemiology and Therapeutic Risk Management, August 22–26, 2015, Boston, Massachusetts, USA. Pharmacoepidemiol Drug Saf. Sep 2015;24:1–587.

[GRADE2004] Grading quality of evidence and strength of recommendations. BMJ. 2004;328(7454):1490.

[Hirsch2013] Hirsch AG, McAlearney AS. Measuring Diabetes Care Performance Using Electronic Health Record Data The Impact of Diabetes Definitions on Performance Measure Outcomes. American Journal of Medical Quality. 2013

[HS2015] Health Search. VIII Report Health Search. https://healthsearch.it/documenti/Archivio/Report/VIIIReport_2013-2014/VIII%20Report%20HS.pdf

[ISTAT2015] Italian National Institute of Statistics. Condizioni di salute e ricorso ai servizi sanitari: informazioni sulla rilevazione. http://www.istat.it/it/archivio/7740

[Landis1977] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.

[OECD2015] OECD. OECD Reviews of Health Care Quality. Italy 2014: Raising Standards. OECD Publishing; 2015.

[Sakshaug 2014] Sakshaug JW, Weir DR, Nicholas LH. Identifying diabetics in Medicare claims and survey data: implications for health services research. BMC Health Serv Res. 2014;14:150.

[TheMatrix] TheMatrix. The tool for easy observational health data management from CSV files. http://thematrix.isti.cnr.it/

[Visca2013] Visca M, Donatini A, Gini R, Federico B, Damiani G, Francesconi P, et al. Group versus single handed primary care: a performance evaluation of the care delivered to chronic patients by Italian GPs. Health Policy. 2013;113(1–2):188–98.

PART II

VALIDITY OF VARIABLES IN MULTI-DATABASE STUDIES

CHAPTER 6

Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research: A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies

Rosa Gini [1,2], Martijn Schuemie [3,4], Jeffrey Brown [5], Patrick Ryan [3,4], Edoardo Vacchi [6], Massimo Coppola [7], Walter Cazzola [6], Preciosa Coloma [2], Roberto Berni [1], Gayo Diallo [8], José Luis Oliveira [9], Paul Avillach [5], Gianluca Trifirò [2], Peter Rijnbeek [2], Mariadonata Bellentani [10], Johan van Der Lei [2], Niek Klazinga [11], Miriam Sturkenboom [2]

1. Agenzia regionale di sanità della Toscana, Osservatorio di epidemiologia; 50141 Florence, Italy
2. Department of Medical Informatics, Erasmus Medical Center; 3015 GJ Rotterdam, The Netherlands
3. Janssen Research & Development, Epidemiology; Titusville, New Jersey, United States
4. Observational Health Data Sciences and Informatics (OHDSI); New York, New York, United States
5. Harvard Medical School, Boston, Massachusetts, United States
6. Università degli Studi di Milano, Dipartimento di Informatica, Milano, Italy
7. Consiglio Nazionale delle Ricerche, Istituto di Scienza e Tecnologie dell'Informazione, Pisa, Italy
8. Université Bordeaux, LESIM - ISPED, Bordeaux, France
9. University of Aveiro, DETI/IEETA, Aveiro, Portugal
10. Agenzia Nazionale per i Servizi Sanitari Regionali; 00100 Rome, Italy
11. Academic Medical Center, University of Amsterdam; 1100 DD Amsterdam, The Netherlands

Published in
Gini et al., eGEMs (Generating Evidence & Methods to improve patient outcomes): Vol. 4: Iss. 1, Article 2. http://repository.edm-forum.org/egems/vol4/iss1/2

ABSTRACT

Introduction

We see increased use of existing observational data in order to achieve fast and transparent production of empirical evidence in health care research. Multiple databases are often used to increase power, to assess rare exposures or outcomes, or to study diverse populations. For privacy and sociological reasons, original data on individual subjects can't be shared, requiring a distributed network approach where data processing is performed prior to data sharing.

Case Descriptions and Variation Among Sites

We created a conceptual framework distinguishing three steps in local data processing: (1) data reorganization into a data structure common across the network; (2) derivation of study variables not present in original data; and (3) application of study design to transform longitudinal data into aggregated data sets for statistical analysis. We applied this framework to four case studies to identify similarities and differences in the United States and Europe: Exploring and Understanding Adverse Drug Reactions by Integrative Mining of Clinical Records and Biomedical Knowledge(EUADR), Observational Medical Outcomes Partnership(OMOP), the Food and Drug Administration's (FDA's) Mini-Sentinel, and the Italian network—the Integration of Content Management Information on the Territory of Patients with Complex Diseases or with Chronic Conditions (MATRICE).

Findings

National networks (OMOP, Mini-Sentinel, MATRICE) all adopted shared procedures for local data reorganization. The multinational EU-ADR network needed locally defined procedures to reorganize its heterogeneous data into a common structure. Derivation of new data elements was centrally defined in all networks but the procedure was not shared in EU-ADR. Application of study design was a common and shared procedure in all the case studies. Computer procedures were embodied in different programming languages, including SAS, R, SQL, Java, and C++.

Conclusion
Using our conceptual framework we found several areas that would benefit from research to identify optimal standards for production of empirical knowledge from existing databases.

INTRODUCTION

Observational studies based on secondary use of existing data collected in the process of health care delivery have the potential to deliver sound evidence quickly enough to support health policy making, which it is often subject to time constraints [1] thus complementing evidence generated by means of primary data collection. However, some epidemiological questions, especially those concerning rare events, rare exposures, and small groups of patients, require more data than is available in any single observational database.[2,3,4] Therefore a growing number of studies use data from networks of databases, sometimes from different countries. Although some of these networks were formed ad hoc for a particular study, several more permanent networks have now been established, where the partners have agreed on an infrastructure and workflow to be reused for different studies.

Privacy regulations and concerns about data ownership and interpretation prevent easy central pooling of original health care data that is now stored in different databases and can be used for secondary purposes.[5] In spite of these barriers several approaches can be used to still employ this data for secondary purposes and pool the results. For example, investigators at each data source can independently create a protocol and execute the study, and estimates are only generated afterward through meta-analysis. A further step is to share the protocol across sites, but asking the local partners to adapt it to their local data and to implement it in their own usual software, to produce local estimates for meta-analysis that are compatible by design. However, most networks now go even further and adopt a *distributed analysis approach*: each database is locally transformed to a representation that is similar across the network, and one single computer program performing the analysis is shared and executed at each site.[4,6]

The need to pool data across different databases is most pronounced in the area of drug safety surveillance.[7] In Europe, the Exploring and Understanding Adverse Drug Reactions by Integrative Mining of Clinical Records and Biomedical Knowledge Project (EU-ADR)[8,9] was initiated in 2008 for investigating the feasibility of signal detection across multiple health care databases. Meanwhile, the United States Food and Drug Administration's (FDA's) Mini-Sentinel Project[10] was developed to support medical product safety monitoring and now includes 18 data partners within a distributed

network. Also in the United States, from 2010 to 2014 the Observational Medical Outcomes Partnership (OMOP)11 performed methodological research on drug safety studies and developed tools and a database network for performing risk identification. Other networks have been developed in other countries, like the Canadian Network for Observational Drug Effect Studies (CNODES) project in Canada and the Asian Pharmacoepidemiology Network (ASPEN) network in Asia.[4] Pharmacoepidemiology is not the only field where the opportunities for combining multiple databases are increasing: in the context of public health or health services research, gathering data from different regions or countries has the added value that different policies can be compared. Mini-Sentinel and EU-ADR are also used to evaluate the impact of regulatory actions. And the Italian network—the Integration of Content Management Information on the Territory of Patients with Complex Diseases or with Chronic Conditions (*Integrazione dei Contenuti Informativi per la Gestione sul Territorio di Pazienti con Patologie Complesse o con Patologie Croniche*)(MATRICE) Project,[12,13] funded by the Italian Ministry of Health— created a distributed network to evaluate the impact of health policies on quality and equity of health care.

We developed a conceptual framework to analyze the process of data management in a network of databases adopting the distributed analysis approach to perform observational studies. We applied the framework to four case studies, and identified similarities and substantial differences.

Purpose and Target of This Study

The purpose of this study was to compare processes that share the same aim but are presently described in separate scientific papers or other documents. Our intent was to find which choices were common among different networks and what the differences were. The comparison findings highlight topics for research. Research should be aimed to further explore if common choices are indeed optimal, and to assess which among the observed differences have an impact on the quality of the processes and on the generated evidence: as such, our findings may be of interest for researchers in medical informatics and methodologists of observational studies. Moreover, the framework and the findings from the comparison provide a unified presentation of strategic choices that are of interest to researchers who are setting or modifying their own networks.

METHODS

Sampling and Data Collection

Some of the paper's authors first conceived of the conceptual framework as an abstraction of the process in place in the European network EU-ADR and in the Italian network MATRICE. They reached out to the authors participating in the United States networks OMOP and Mini-Sentinel, to compare networks of different continents. Data collection was performed via document (scientific papers and websites) analysis and interviews with coauthors. The manuscript was reviewed by all the authors.

The Four Networks

The EU-ADR Project was funded by the European Commission under Framework Programme [7] (FP7) and ran from 2008 to 2012 with the aim of producing a computerized integrated system for the early detection of drug safety signals. The project used data from eight databases from four European countries (Denmark, Italy, the Netherlands, and the United Kingdom) covering a population of about 20 million individuals overall with almost 60 million person-years (PYs) of follow-up.[3] Subsequently, the EU-ADR workflow has been further improved and applied in several collaborative drug-safety studies concerning NSAIDs (SOS),[14,15] pandemic influenza vaccine (VAESCO),[16] the arrhythmogenic potential of drugs (ARITMO),[17] and hypoglycemic drugs (SAFEGUARD).[18] The subjects of the studies performed in this network include methodology[19,20,21,22] drug utilization, disease incidence,[23] signal detection,[24] testing,[25,26] filtering,[27] and substantiation.[28] The workflow is currently being extended in the European Medical Information Framework (EMIF) project.[29]

The United States FDA's Mini-Sentinel program[30] began in 2008 and has created a distributed data network of 18 data partners covering a population of over 150 million persons and 380 million PYs in the United States.[6] Mini-Sentinel was structured to produce both fast, standardized replies to specific queries (called Rapid Response queries) and studies based on ad hoc developed protocol (i.e., Protocol-based Assessments). Hundreds of Rapid Response queries are executed each year, and 14 Protocol-based Assessments have been completed or are underway. Network activities cover a broad range of topics including drug utilization, disease burden, the impact of regulatory policies,

and the comparative safety of medical products.[31] At the same time, several studies focusing on methodology have been completed.[32,33,34,35,36,37] In 2015 the Mini- Sentinel pilot transitioned to the Sentinel system that is become part of the FDA's regulatory framework.

OMOP was a public-private partnership that ran from 2010 to 2014 and was part of the Innovation in Medical Evidence Development and Surveillance (IMEDS) program of the Reagan-Udall Foundation for the FDA. Its goal was to help determine best practices for use of observational health care data. OMOP currently maintained five commercial databases covering 164.[9] million persons in its own central venue, and its data partner network included six other databases covering an additional 105 million persons.[8,38] The network was used to develop tools for performing observational studies in a database network, including the OMOP Common Data Model (CDM),[39] the OMOP Vocabulary,[40] and tools for assessing data quality,[41] as well as research into the development and evaluation of methods for drug-associated risk identification.[42] In 2014 the OMOP research team launched the Observational Health Data Sciences and Informatics (OHDSI) (pronounced "Odyssey") program[43] which is currently continuing the activity of OMOP.

The MATRICE project was funded by the Italian Ministry of Health and ran from 2011 to 2014 under the coordination of the Italian National Agency for Regional Health Services to measure quality of health care for chronic diseases. MATRICE developed a distributed network infrastructure specific to local and regional Italian administrative databases and is rapidly growing to include participants beyond the project. Currently, it covers a population of about 9 million subjects living in some of the Local Health Authorities in 9 of the 21 regional health care systems in the country. Studies completed so far using data from this network were aimed at evaluating the quality and equity of primary care, the impact of policies in this field [44,45,46,47] and methodological challenges of such studies.[48,49] The network currently participates in several studies funded by the Italian Ministry of Health.

Conceptual Framework

Figure 1 depicts our conceptual framework, showing a workflow consisting of data sets (D1, D2, D3, and D4) and transformation processes (T1, T2, and T3). The conceptual framework does not contain recommendations in itself: it is just a conceptual abstraction of the logical sequence of steps needed to

Box 1. Definition of the Conceptual Framework

DATA SETS AND DATA TRANSFORMATIONS

D1 (original databases: DBs) is a collection of data sources controlled by a single organization that has procedures in place to link them with each other at the individual level, thus creating a single data pool on the same subjects. The term "DB" refers to an organization that has access to the data.
T1 (data reorganization) is a data modeling step: transformation from the locally defined data repository into a global (common) schema with standardized variable and attribute names, without loss of information. Simple one-on-one recoding is performed as well, such as making data formats and coding of attributes (e.g., gender) identical. T1 is specific per DB but independent of the specific study.
D2 (global schema, GS) is a general database schema that contains all the attributes thatare necessary to answer a realm of study questions ("use cases") that are of general interest to the network, such as incidence of disease, drug utilization, or association studies. D2 has a defined set of table names, attribute names, and formats. D2 plays the same role as a GS of a data integration system.50 Therefore, a set of correspondences are defined between this schema and the D1. Note that (1) these correspondences may not be complete for all databases: for instance, if a D1 does not have information about primary care diagnoses, these attributes will remain empty in the D2; and (2) some attributes (typically, diagnoses or drugs) might have different coding for different DBs in the network.
T2 (data derivation) is the step where novel meaning is obtained from D2 by means of an explicit manipulation and combination of D2 data. These manipulations are necessary when a study variable is not among those collected by one of the DBs in the network, and must therefore be represented, by proxy, as a combination of whatever pertinent information is available. When the study variable is a disease, this process is referred to in the literature as disease phenotyping.51 T2 is often specific per DB, as it depends on the information that was originally collected, and is often specific per study, although conceivably past data derivations could be reused in new studies. As an example of T2, if the presence of diabetes in study subjects needs to be assessed, DBs collecting data from primary care can identify the information from a general practitioner's (GP's) diagnosis, whereas claims databases without clinical data from primary care may use dispensing of antidiabetic drugs as proxy, and combinations may also be possible.
D3 (derived data) are the data sets derived in T2, each containing one or more study-specific variables. Derived data may be occurrence of a disease, or other information like the duration of exposure to a specific drug. For instance a drug safety study has three basic types of derived data: the outcome of interest (often sudden occurrence of a condition), the exposure (a sequence of drug utilization episodes), and presence in the study cohort, with beginning and end dates of follow-up. While the tables forD2 contain multiple, longitudinal observations per subject, each generated during an encounter and each containing multiple codes, D3 contains as many observations per subject as requested by the study design (often one single observation). Original data (as modeled in D2) is therefore "rolled up"during T2 to create in D3 the best possible approximation of the variables needed in the specific study.
T3 (study design application) is data transformation for a specific analytic: based on the protocol of a study with specific design (application of inclusion and exclusion criteria, selection of exposure windows, propensity and disease score estimation, control selection, matching). T3 produces the data sets for statistical analysis. Within this transformation data may be de-identified and aggregated to various levels. T3 is specific to the study, butis the same across participant DBs.
D4 (data sets for analysis) is the result of T3. D4s from all the partners in the network are similar. Based on the level of sharing that is allowed, D4 may stay local at the database custodian or be pooled in a central repository. In both situations, statistical analysis on D4 follows and produces estimates to be interpreted.

QUALITY

1. **Process verification**: assuring quality, transparency and reproducibility of the stepwise data extraction process, e.g., common standard process documentation, process automatization with common use of dedicated software, and parallel programming; and
2. **Outcome verification:** checking intermediate and final output against standards, including the following:
• Benchmarking of D3 (derived data) against external data (e.g., determining whether observed disease rates are in line with those reported in literature);
• Benchmarking of D3 within the network (comparison of DB-specific output to assess homogeneity);
• Validation of D3 using a gold standard (e.g., chart review) to assess performance of data derivation (e.g., positive predictive value); and
• Validation of D4 using expected results (i.e., using a reference set of known causal or non causal associations).

perform studies in a network. Figure 2 describes each step in detail. During a typical study, data transformation T2 and T3 might be performed iteratively: if additional analyses are required to shed light on preliminary results, then T3 or both T2 and T3 can be repeated and new D4 can be produced to undergo statistical analysis. In some studies T2 (data derivation) may not be performed, if data needed for the study are all contained in the original data.

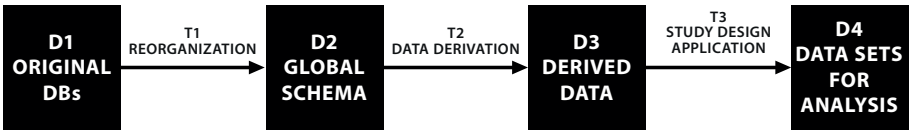


Figure 1. Flowchart of the Data Transformation Process Occurring Locally in a Study Collecting Data from a Network of Databases. D1, D2, D3, and D4 represent data sets; T1, T2, and T3 represent data transformations.

To ensure that T1–T3 are valid, both in terms of how well the transformation reflects the original data and of whether it achieves the aim of the transformation, quality control processes need to be in place. In Box 1 process and outcome verification steps are highlighted.

To illustrate the steps of the workflow, an example from the MATRICE network is shown in Box 2.

Box 2. An Example of Data Management in the MATRICE Network

The Italian National Agency for Regional Health Services promoted a study to assess whether regional Italian administrative databases can be used to measure whether patients with Chronic Obstructive Pulmonary Disease (COPD) are treated with recommended therapies. The study objective was to establish whether different cohorts, defined with different case-identification strategies, resulted in consistent estimates of therapy adherence. The MATRICE network was used for this study.
Five regions were involved in the study. In each Italian region several tables of administrative data are collected with content regulated by national law, in particular the following: the list of residents (citizens and regular migrants) entitled to receive health care; hospital discharge records, with six diagnosis codes; exemptions from copayments for health care; and drug prescriptions. In each region participating in the study, a copy of the four tables (D1) was stored, with different data models and format. The MATRICE network has established a specific data model for the above mentioned four tables(list of residents; hospital discharge records; exemptions from copayments for health care; and drug prescriptions), and the format is flat comma-separated files (D2). Two of the regions had already participated in a previous study of the MATRICE network, so T1 had already been performed. In the other three regions, the format D2 was explained to a local expert by means of structured documents and a teleconference, a common software named TheMatrix was installed (see "T1: reorganization" in the "FINDINGS" Section below), and T1 was performed by the local expert and was checked with standard procedures embedded in the software. The study protocol had defined several variables to be extracted or derived: gender, presence in the region at index date, age at index date, presence of a COPD diagnosis in the 1–5 years before index date, presence of some patterns of utilization of respiratory drugs in the 1–3 years before index date, and adherence to recommended therapies during follow-up. D3 was composed ofa group of data sets, one per derived variable, each with a single observation per subject. Since in MATRICE all the participating data partners share the same data content (see "D1: original DBs" in the "FINDINGS" Section below), the transformation T2 was uniform across data partners. T2 was therefore embedded by the principal investigator in a single ad hoc procedure of the software TheMatrix, shared with the local partners and executed locally. The data set D4 was designed in the protocol to be the aggregated data set that counted the frequency of each combination of the variables in D3. The transformation T3 was embedded by the principal investigator in another ad hoc procedure of the software TheMatrix, shared with the local partners and executed locally. The D4s produced by the five regions were shared with the principal investigator, who executed the statistical analysis of the pooled data set using the statistical software Stata 13.1.

FINDINGS

We describe and compare T1–T3 and D1–D4 in the four networks.

D1 (Original DBs)

We use "DB" to refer to an organization that has access to the data. Table 1 summarizes the DBs participating in the four networks. For each network a column represents a combination of data sources that are linked in at least one database. We classified data sources according to their provenance, and we indicated the data items available in the DB from that data source. If more than one DB in a network share the same combination, only one column is shown: the number of columns fora network in Table 1 is therefore a measure of heterogeneity of the DBs participating in the network. MATRICE has a single combination (M1), EU-ADR has seven (EU1–EU7), Mini-Sentinel has three (MS1–MS3), and OMOP has four (O1–O4)

Table 1. Description of the D1 (original DBs) databases in terms of provenance and data items collected from each data source

Provenance of Data Source		Combinations of Data Sources and Data Items Available in the DBs of the Network														
		MATRICE	EU-ADR							Mini-Sentinel			OMOP			
		MA1	EU1	EU2	EU3	EU4	EU5	EU6	EU7	MS1	MS2	MS3	O1	O2	O3	O4
Primary care	Administrative data									Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc	
	Clinical data						Dx	Dx Rx Text Refspec Refinpat Vac	Dx Rx Text Refspec Refinpat Res	Dx Proc						Dx Rx
Secondary care	Administrative data	Spec Proc	Spec Proc	Spec Proc	Dx Proc					Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc	
	Clinical data									Dx Proc						Dx Rx Text
Inpatient care	Administrative data	Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc				Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc	Dx Proc	
	Clinical data									Dx Proc						
Enrollment into the data collection		Geo	Geo	Geo	Geo	Geo	Charge	Charge	Charge	Elig	Elig	Elig	Elig	Elig	Elig	Elig
Pharmacies		Rx	Rx	Rx	Rx	Rx				Rx	Rx	Rx	Rx	Rx	Rx	
Registry of disease-specific exemptions from health carecopayment		Dx	Dx	Dx												
Death registry				Dx	Dx					Dx	Dx	Dx				
Vaccination registry										Vac	Vac					
Laboratory		Lab	Lab	Lab	Res	Res		Res		Res	Res			Res		

Notes: If more than one database in a network has access to the same combination of data, they are represented by a single column. Data items—Dx: diagnostic codes; Proc: procedure codes; Rx: prescriptions or dispensings of drugs; Spec: specialty of secondary care encounters; Refsec: Referrals from secondary care; Refinpat: Referrals from inpatient care; Text: Notes in free text; Lab: Labels of laboratory tests; Res:Laboratory test results;Vac: Vaccines; Geo: Presence in a geographical area; Charge: Being in assisted by a GP; Elig: Satisfying eligibility criteria for an insurance company or health plan.

Differences and Similarities

First, in the two United States–based networks (OMOP and Mini-Sentinel) almost all databases (O1–O3 and MS1–MS3) obtain administrative information from primary, secondary, and inpatient care, while in both European networks (EU-ADR and MATRICE) each database lacks at least one setting. Second, EU-ADR pools data from the most heterogeneous databases: the eight databases showed seven different combinations. Third, in Italy, although administrative information from secondary care (such as specialty of the physician visiting the patient) is available, it does not contain diagnostic codes (M1 and EU1–EU2). Fourth, access to laboratory test results is rare among databases in all networks. Fifth, in all but one United States database, enrollment of subjects in the data collection is due to the eligibility criteria for social insurance or an insurance company, while in Europe criteria include geographical residence or being listed with a GP. Sixth, only in EU-ADR and Mini-Sentinel are death and immunization registries available. Finally, only Mini-Sentinel involves partners collecting information from both clinical and administrative data sources. This is achieved by integrated delivery systems that operate medical facilities from which they collect electronic health care records data. In addition, all the

Box 3. Short Illustration of the Differences in Original Data

In 2005, Irina, age 36, developed gestational diabetes during her second pregnancy, which was diagnosed by her gynecologist and treated with insulin prescribed by her GP. Irina gave birth to Louise in a hospital, and had her vaccinated against tetanus and diphtheria when the baby was six months old. The following year Irina's father Mario, age 67 and a smoker with a history of coronary heart disease, moved to the region where Irina lived. In 2007, Mario was diagnosed with diabetes by his GP, who was also his daughter's GP. After trying for a while to cope with his condition only through following a new diet, he started taking antidiabetic drugs in 2008. In 2010 he had severe angina and was admitted to the hospital for a few days. In 2013 Mario died in his sleep, and his death certificate indicated that the cause of death was myocardial infarction.

If Irina, Louise, and Mario were part of the database population of the four networks, the image of the story would be different. For databases lacking diagnosis from primary or secondary care, like M1 or EU1–EU4, Irina's beginning to take insulin could be misinterpreted as an occurrence of diabetes, even though a complex algorithm using hospital admittance for delivery or the ending of insulin prescriptions could effectively avoid misclassification. Louise's vaccine would be detected by MS1, MS2, and MS3. When Mario moved to Irina's region and entered the database population, only databases collecting clinical history from primary care—like EU6, EU7, MS1, and O4—could have detected that he was the father of Irina and was a smoker. While the history of coronary heart disease could also be deduced from the same databases or clinical notes of a cardiologist in MS1, the presence of the disease may be inferred from drug utilization data in all the databases, and angina precisely in 2010 in databases with diagnoses from inpatient care (MA1, EU1–EU4, all MS, and O1–O3). Diabetes would be detected in 2007 from primary care diagnosis in EU5–EU7 and all the United States databases, and in 2008 only from drug utilization in the others. Occurrence of myocardial infarction would be detected only by EU2, EU3, and all the MS databases.

partners of Mini-Sentinel and some partners of the other networks can access full-text medical records for chart validation for their population.
Box 3 is a fictional example of the impact of the differences in D1 on the information captured from a patient history.

T1 (Reorganization)
In Table 2, T1 is compared across case studies.

Differences and Similarities

Besides local storage, in OMOP some databases also allow creating a central and cloud-based copy of the transformed data. In MATRICE and Mini-Sentinel, all original databases used the same coding systems, while in OMOP participating databases used different coding systems and even unstructured free text in different languages, in EU-ADR.
Different strategies were adopted to transform the original data into a common data set: in EU-ADR, the transformation T1 was used only in internal discussions to define T2, and data sets in the common data model were never created. In MATRICE, standard procedures for T1 are in place, and results are evaluated by local partners. In Mini- Sentinel, data is transformed to a general, common data model and is updated frequently; and checks for data completeness and consistency with the data model are Standard Operating Procedures (SOP) executed as part of each transformation and approval process.[52] OMOP recoded all data to a single system during T1, independently of a study question, the transformation in T1 is evaluated by first generating descriptive statistics of all elements in D2 using a tool called Observational Source Characteristics Report (OSCAR), and by subsequently performing internal and external comparison of these statistics using a tool called Generalized Review of OSCAR Unified Checking (GROUCH). Both in OMOP and Mini-Sentinel, a formal Extraction, Transformation, and Loading (ETL) document is created as part of development and implementation of the data model. In MATRICE the transformation is executed via ad hoc software, called TheMatrix,[53] whose configuration is stored in a text file.

Table 2. Comparison with respect to T1, D2, T2

T1 (data reorganization)			
Network	**Recoding**	**Quality: data completeness**	**Quality: documentation**
EU-ADR	Does not require mapping to external standard: original coding or free text is maintained.	Demanded to local partners, no formal procedure	No formal documentation
Mini-Sentinel	Source data are homogeneous in coding systems.	Local report on specific issues,plus feedback from standard programs checking for completeness and consistency	Data model, data elements, and guiding principles approved by partners. ETL formal document, ad hoc per DB
OMOP	Source data standardized to common vocabulary by domain: Drug (RxNorm), Condition (SNOMED), Labs (LOINC)	Formal procedures: OSCAR and GROUCH tools	ETL formal document, ad hoc per DB
MATRICE	Source data are homogeneous in coding systems	Formal procedures checking data completeness	Local configuration of the TheMatrix software (text file)

D2 (Global schema)				
Network	**Table Names based on**	**AttributeNames based on**	**Every CDM table has a view in every DB**	**Attributes are coded uniformly across DBs**
EU-ADR	Reason/setting of data recording	Clinical contents	N	N
Mini-Sentinel	Clinical content and data source (diagnosis, procedures, encounters, lab results); or reason/setting (outpatient pharmacy, death, enrollment)	Reason/setting of data recording for diagnosis and similar, clinical contents for pharmacy and death	N	Y
OMOP	Clinical content	Reason/setting of data recording	Y	Y
MATRICE	Reason/setting of data recording	Clinical contents	Y	Y

T2 (data derivation)				
Network	**Logic**	**Single definition per derived data**	**Quality: process control**	**Quality: validation**
EU-ADR	DB-specific algorithms, harmonized through a formal negotiation process	Y	No common procedures were implemented, although logic of local procedures was shared.	Internal incidence rates comparison, literature comparison, some validation with external gold standard (PPV)

Table 2. Continued

	T2 (data derivation)			
Network	**Logic**	**Single definition per derived data**	**Quality: process control**	**Quality: validation**
Mini-Sentinel	The same algorithm was used across all DBs.	Y	Shared SAS script	Systematic review of previously published validation studies, expert clinical, data, and epidemiologic guidance, medical chart review for PPV, and assessment of difference in dates
OMOP	Multiple alternative algorithms were adopted to derive the same data; some were DB-specific.	N	Shared parameterized SQL queries stored in common procedure (RICO)	Internal prevalence rates comparison, no external validation performed
MATRICE	Multiple algorithms were explored, decision was taken by means of a validation study.	Y	Shared script in a scripting language developed ad hoc (TheMatrix)	Validation of algorithms with external gold standard: sensitivity, specificity, PPV, NPV

D2 (Global Schema)

In Table 2, D2 is compared across case studies.

Differences and Similarities

The main difference we observed in the evaluation of the data models was the way two main characteristics of an encounter were captured: the setting where the health care was administered (e.g., general practice, inpatient care, laboratory) and the medical content of the encounter (e.g., diagnosis, procedure, laboratory test). One possibility was that information was grouped in tables according to the setting (e.g., a table for hospital admissions, another for laboratory tests) and facts were recorded as attributes. The alternative was that encounters were grouped in tables defined by medical content (e.g., a table for diagnoses, a table for procedures) and the care setting was recorded as an attribute. EU-ADR and MATRICE adopted the first approach, OMOP adopted the second, and Mini-Sentinel adopted a combination of the two approaches—death and pharmacy dispensations were organized in the first way and other information was organized in the second.

Table 3. Comparison with respect to T3 and D4

T3 (application of study design)					
Network	**Local partners execute shared procedure**	**Common among DBs**	**Scores estimation**	**Specific software**	**Programming language**
EU-ADR	Y	*Y*	N	Jerboa	Java andJerboa scripting languages
Mini-Sentinel	Y	Y	Y	Modular programs and macros; PopMedNet	SQL, SAS, Java, R,
OMOP	Y	Y	Y	-	SQL, SAS, R, C, Java
MATRICE	Y	Y	N	TheMatrix	Java and TheMatrix scripting languages

D4 (data sets for analysis)			
Network	**Type**	**Format**	**Quality: study results validation**
EU-ADR	Intermediate files that can be shared among partners, analysis will follow	CSV	Drug safety methodology: comparison of observed drug-event associations with previously classified true and false causal associations; impact on this of different definitions of the derived data
Mini-Sentinel	Level of granularity of data set depends on study needs; always transfer minimum necessary. Some analyses transfer aggregate data, some use highlysummarized patient-level data. Intermediate files saved locally by data partners.	CSV, SAS datafiles, HTML	To test code known associations are used. Rapid Response queries include data characterization and are reviewed manually by a data expert and an epidemiologist. Results are also reviewed by data partners. Protocol-based assessments might include chart reviews.
OMOP	Final estimates, intermediate files are discarded.	CSV, SAS data files, SQL tables	Drug safety methodology: comparison of observed drug-event associations with previously classified true and false causal associations; impact on this of different definitions of the derived data; estimate of residual bias per event by means of known noncausal associations.
MATRICE	Intermediate files to be used for analysis or report generation	CSV	Results are reviewed by data partners for comparison with similar analysis performed independently.

T2 (Data Derivation for Specific Studies)

In Table 2, T2 is compared across case studies.

Differences and Similarities

In EU-ADR each data custodian executed its algorithm with its own usual extraction tool to derive simple input files for a specific study, while execution was performed with common software on the GS in the other networks.

OMOP and Mini-Sentinel adopted shared SQL and SAS code, respectively. In MATRICE an ad hoc scripting language was designed and a compiler (a computer program that transforms source code written in a programming language into another) from this language toward the Java virtual machine was developed; extraction in a shared code was then executed locally. Since OMOP focused on methods development, it often used multiple algorithms for data derivation, to study the impact of the differences. In MATRICE, ongoing validation studies test several algorithms, but the plan is to use a single best definition per study in the end.

In EU-ADR, to overcome the heterogeneity across terminologies, a shared semantic foundation was built by using Unified Medical Language System (UMLS) concepts to define events5. Then, the definitive choice of algorithms was obtained through an iterative negotiation between databases: DBs with similar structures were invited to query the same tables and fields.54 In Mini-Sentinel, algorithms are developed (or reused) for specific analyses and applied at the time of analysis; the result of those algorithms is not stored in the database, but analytic files for each assessment are retained locally.

As for validation of the event resulting from data derivation, all the networks compared incidence or prevalence rates among databases as a tool to assess consistency. OMOP did not routinely compare with external standards nor with the literature. The other networks performed either population-based external validation to estimate all validity indices (MATRICE) or external validation of a random sample of automatically detected events to estimate positive predictive value (EU-ADR, Mini-Sentinel).

T3 (Study Design Application)

In Table 3, T3 is compared across case studies.

Differences and Similarities

During steps T1 and T2, local partners in some of the networks were asked to implement the processes that had been agreed upon in their own local procedures; moreover the procedures were not shared. In step T3 (study design application), data transformation into analytical data sets was performed in all four networks using shared and common software. In Mini- Sentinel and OMOP, statistical analysis was needed in T3 to estimate propensity and disease

scores, while in the studies implemented in the other networks only simpler tasks were needed: linkage between different tables, time splitting, random selection, matching, de-identification, and aggregation. The software Jerboa was developed and used by EU-ADR to execute T3. The software TheMatrix developed by MATRICE executes both T2 and T3: a Domain Specific Language (DSL) was designed and developed for this purpose. DSLs are computer programming languages whose features and expressiveness are restricted and designed ad hoc to fit a given field of application. They target a narrower set of programs than general-purpose languages like Java, but in exchange they provide a higher level of abstraction and can be programmed directly by domain experts rather than computer programmers.[55] In MATRICE, a DSL generating tool called Neverlang was used to develop the language,[16,56] and scripts in the language were generated by domain experts. Mini-Sentinel and OMOP both used existing software (SQL, SAS, C, Java and R).

D4 (Data Sets for Analysis)
In Table 3, D4 is compared across case studies.

Differences and Similarities

In OMOP only final estimates were shared, while in the other networks integrated data sets were shared to be pooled before statistical analysis. EU-ADR and OMOP both adopted a similar validation strategy for their methodological studies in drug safety, which implicitly validated the whole sequence of data transformations at once: a set of positive controls (known adverse drug reactions) and negative controls (drug-outcome pairs that are believed to have no causal relationship) was created. The quality of each method of analysis was assessed by measuring its discriminating power, i.e., the ability of telling positive from negative controls.

DISCUSSION

In this paper we introduce a conceptual framework to analyze the data management process of a network performing distributed analyses. By applying the framework to four case studies we identify similarities and substantial

differences. With this as the foundation, we highlight areas that need further research to identify optimal strategies.

Differences in Original Databases (of DBs) Have Huge Consequences

The differences observed in the four networks when comparing the original databases (D1) are huge. Understanding such differences is a challenge in itself, as terminology describing health data sources is not shared across countries. [57] The three national networks (MATRICE, OMOP, and Mini-Sentinel) were much more homogeneous than EU-ADR. Since we expect that networks will continue to grow and new DBs will be different from existing DBs, the problems that EU-ADR encountered could indicate challenges other networks will face in the future if the geographical area is extended. United States databases often have in- and outpatient diagnoses, whereas these are rarely all captured in European administrative databases. In contrast, in Europe general practice databases are very rich since in many countries GPs have a gatekeeper function, that is, nonemergency health care can be accessed free of charge only upon the prescription of a GP. Death registries are infrequently part of the data sources available to databases, and this hampers detection of conditions, like acute myocardial infarction or stroke, which may cause death before the patient can reach a health care facility. Due to the differences in available information in the different databases, various strategies need to be used in order to have a comprehensive data derivation of study variables, e.g., in the absence of outpatient diagnostic data, drug utilization or laboratory values may be used to identify certain conditions.

Differences in global schemas are not substantial

Differences in the GS (D2) between the networks exist but are not substantial, as each GS can be mapped into another, except for those data items that are specifically collected in a single network (for instance, exemptions from copayment, which are documented only in Italian DBs). It would be very valuable, however, to explicitly create such a mapping, as this would make it possible to run existing software procedures embodying T2 and T3 independently of the network: this happened, for instance, in a study replicating—in the EU-ADR network—results from the OMOP network.[58] One area of research should be the impact of different formats of GSs on study outcomes.

Different Approaches to Terminology Mapping

In two networks (OMOP and EU-ADR), different disease and drug coding systems needed to be managed. In OMOP the differences were addressed by mapping to homogeneous coding systems during T1, although the original codes were not discarded but were also included in D2. In EU-ADR, mapping was not conducted in T1, therefore all mapping was performed during T2 and only for study-specific conditions. Due to the large differences in the granularity and type of coding schemes, in European databases mapping was very time-consuming—yet this was necessary to obtain consensus across data custodians and investigators[6]—and is progressively growing a shared library. The impact of different mapping strategies, and whether mapping should be done at all versus addressed in the analytic phase, should be investigated.

Sharing Aggregated Data Sets Versus Sharing Estimates

If network partners can share aggregated data sets in D4, the investigators maintain freedom to perform some subset and sensitivity analysis that were not strictly foreseen in the protocol without performing a new round of transformation. Sharing aggregated data would allow different levels of pooling and potentially more power with respect to meta-analysis, although previous research shows no improved performance of one approach over the other. [14,59,60,61] Given the privacy related issues around data sharing, it should be investigated when different levels of sharing may be indicated.

Software Tools, Professional Skills, and Information Technology

Software tools used during the transformation process differed across case studies. This had implications for the type of professional skills needed to perform studies in the network as well as the readability of the programs for other investigators. In principle, all data transformations must be documented to allow investigators to correctly interpret study results and to understand study limitations and strengths. OMOP and Mini-Sentinel have complete websites where information is stored and can be openly accessed, while EU-ADR and MATRICE rely mainly on scientific papers and reports, a less efficient way of storing information. How to develop transparent programs and how to store and share the corresponding complex body of information to make it easily available to investigators is also a relevant research topic.

Validation

Validation of derived data is an imperative condition to produce good epidemiological estimates,[62] and this is even truer when heterogeneous databases participate ina network. Indeed, regularizing the process of creating research data sets from secondary data sets, although necessary, is not enough to ensure high data quality; and validation can quantify how much derived data fail in correctly identifying the study variables—failure that can differ across data partners.

In MATRICE—as data from primary care is lacking and information from secondary care is sparse— deriving chronic conditions, the primary focus of the network, is cumbersome. This is why MATRICE is leading a population-based validation study using diagnosis from a sample of GPs as a gold standard. In Mini-Sentinel a model for a typical validation study was developed13 and implemented for some events, in particular acute myocardial infarction.[14] EU-ADR adopted a similar study design in some validation studies.[7,8] Only positive predictive value could be estimated from the study design adopted in the two networks. A similar study was performed on an occasional basis in OMOP.[63] In order to estimate sensitivity, access to a populationbased data source would be required, which is more complex than accessing clinical charts of selected candidate events. However, in the specific case of acute myocardial infarction, death registries are estimated to add from 15 percent to 25 percent of cases to inpatient data where both data sources are available.[6] Therefore misclassification of non-cases, in principle, could have a relevant impact on study results, especially in older subpopulations. In EUADR it was observed that improving the positive predictive values of the outcome definition had a very small impact on estimates of additional risk of upper gastrointestinal bleeding in users of four drugs,[7,64] and in OMOP methodological studies varying the definition of several outcomes had little impact on system performance overall,65 thus suggesting that outcome misclassification may not be a paramount concern when studying the safety of short exposure to drugs. This area has generated some research66 and deserves further study. The only attempt to automatically incorporate the result of a quality procedure in the interpretation of study results was performed in OMOP: the association with an outcome observed in a set of drugs that are a priori known not to cause the outcome was computed and applied as an estimate of overall bias in the association of any drug with the same outcome.[67]

Designing and developing a framework that allows for automatically incorporating validity indices in study design and analysis would be a useful followup for the effort invested in validation.

Epistemological Framework of Reference

Unlike in the other steps, in T3 there was a very similar approach in the four networks: there is a uniform attempt to make study designs clearly specified and reusable across studies. This was achieved in all four networks by embedding this step into shared software, where the same procedure was executed across all data sites. It could be argued that complexity arising from the network setting forces investigators to specify—right from the study design stage —every detail of data management and analysis, embedded in a sequence of computer instructions. A priori specification of the detail of the experiment is at the epistemological core of the experimental method, as it ensures falsifiability.[68] From this point of view, the intricacies of the network settings force investigators to do the right thing. Computer engineers have joined pharmacoepidemiologists and other populationbased health scientists in supporting this effort, not just because computer programming is needed, but also and most of all, because a novel, more formal process must be streamlined and stabilized before investigators take control again of the new level of complexity.

Limitations

The conceptual framework was useful to interpret similarities and differences among the four networks, which are heterogeneous for geographical coverage and purpose. However the choice of the sample of four was nonsystematic, therefore the framework may prove insufficient to include other networks in the comparison. Data processing in networks of databases may suffer from subtle challenges: privacy laws may enable patients to opt out of sharing information based on some encounters only (for instance, for mental health issues); some databases may collect information from smaller health care providers, whose information is not effectively shared in digital form; regional or national differences in privacy regulations may affect differentially the partners of a network. We did not investigate how the four networks faced such challenges.

CONCLUSION

We proposed a conceptual framework to analyze the data management process involved in observational studies taking place in distributed networks of databases. The framework was applied to four case studies to identify similarities and differences. Several research questions were highlighted by this comparison, including interoperability among the available GSs, optimization of data harmonization, use of validity indices in study design and statistical analysis, development of an information infrastructure to support investigators in accessing details of data transformation, and optimal level of programming skills needed to manage the process. Medical informatics is called on to support transparency, and quick and sound application of the experimental method to the production of empirical knowledge.

REFERENCES

1. Salmon D, Yih WK, Lee G, Rosofsky R, Brown J, Vannice K, et al. Success of program linking data sources to monitor H1N1 vaccine safety points to potential for even broader safety surveillance. Health Aff (Millwood). Nov 2012;31(11):2518–27.
2. Hernán MA, Savitz DA. From "Big Epidemiology" to "Colossal Epidemiology". Epidemiology. May2013;24(3):344–5. 3. Toh S, Platt R. Is Size the Next Big Thing in Epidemiology? Epidemiology. May 2013;24(3):349–51.
4. Trifirò G, Coloma PM, Rijnbeek PR, Romio S, Mosseveld B, Weibel D, et al. Combining multiple healthcare databases for post-marketing drug and vaccine safety surveillance: why and how? J Intern Med. 1 Mar 2014.
5. McGraw D, Rosati K, Evans B. A policy framework for public health uses of electronic health data. Pharmacoepidemiology and Drug Safety. Jan2012;21:18–22.
6. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care. Jun 2010;48(6 Suppl):S45–51.
7. Valkhoff VE, Coloma PM, Masclee GMC, Gini R, Innocenti F, Lapi F, et al. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. Journal of Clinical Epidemiology. Aug 2014;67(8):921–31.
8. Avillach P, Mougin F, Joubert M, Thiessard F, Pariente A, Dufour J-C, et al. A semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European eu-ADR project. Stud Health Technol Inform. 2009;150:190–4.
9. Coloma PM, Schuemie MJ, Trifirò G, Gini R, Herings R, Hippisley-Cox J, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol Drug Saf. Jan 2011;20(1):1–11.
10. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiology and Drug Safety. Jan 2012;21:1–8.
11. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership. Ann Intern Med. 11 Feb 2010;153(9):600–6.
12. Agenzia nazionale per i Servizi Sanitari Regionali. Programma Mattoni del SSN - Progetto MATRICE. http://www.agenas.it/ images/agenas/In%20primo%20piano/Matrice/Progetto_ MATRICE_Scheda_informativa.pdf. Accessed September 2015 [Italian]
13. Agenzia regionale di sanità della Toscana. Data integration for chronic diseases management in outpatient settings (MATRICE Project). https://www.ars.toscana.it/en/project/ chronic-diseases/2460-matrice-project.html. Accessed September 2015
14. Cordis. Safety Of non-Steroidal anti-inflammatory drugs. http://cordis.europa.eu/result/ rcn/54210_en.html. Accessed September 2015 15. Valkhoff VE, Schade R, 't Jong GW, Romio S, Schuemie MJ, Arfe A, et al. Population-based analysis of non-steroidal anti-inflammatory drug use among children in four European countries in the SOS project: what size of data platforms and which study designs do we need to assess safety issues? BMC Pediatr. 2013;13:192.
16. Vaccine Adverse Event Surveillance & Communication. https:// brightoncollaboration.org/vaesco. html. Accessed September 2015
17. Cordis. Arrhythmogenic Potential of Drugs. http://cordis. europa.eu/project/rcn/94061_en.html. Accessed September 2015.
18. Safety Evaluation of Adverse Reactions in Diabetes. www.safeguard-diabetes.org. Accessed September 2015
19. Avillach P, Joubert M, Thiessard F, Trifirò G, Dufour J-C, Pariente A, et al. Design and evaluation of a semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European EUADR project. Stud Health Technol Inform.2010;160(Pt 2):1085– 9.

20. Coloma PM, Valkhoff VE, Mazzaglia G, Nielsson MS, Pedersen L, Molokhia M, et al. Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems: a validation study in three European countries. BMJ Open. Jan2013;3(6).
21. Schuemie MJ, Coloma PM, Straatman H, Herings RMC, Trifirò G, Matthews JN, et al. Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. Med Care. Oct 2012;50(10):890–7.
22. Trifirò G, Patadia V, Schuemie MJ, Coloma PM, Gini R, Herings R, et al. EU-ADR healthcare database network vs. spontaneous reporting system database: preliminary comparison of signal detection. Stud Health Technol Inform. 2011;166:25–30.
23. Wijnans L, Lecomte C, de Vries C, Weibel D, Sammon C, Hviid A, et al. The incidence of narcolepsy in Europe: before, during, and after the influenza A(H1N1)pdm09 pandemic and vaccination campaigns. Vaccine. 2013;31(8):1246–54.
24. Coloma PM, Schuemie MJ, Trifirò G, Furlong L, van Mulligen E, Bauer-Mehren A, et al. Drug-Induced Acute Myocardial Infarction: Identifying 'Prime Suspects' from Electronic Healthcare Records-Based Surveillance System. PLoS ONE. Aug 2013;8(8):e72148.
25. Dieleman J, Romio S, Johansen K, Weibel D, Bonhoeffer J, Sturkenboom M, et al. Guillain-Barre syndrome and adjuvanted pandemic influenza A (H1N1) 2009 vaccine: multinational casecontrol study in Europe. BMJ. 2011;343:d3908.
26. Romio S, Weibel D, Dieleman JP, Olberg HK, de Vries CS, Sammon C, et al. Guillain-Barré Syndrome and Adjuvanted Pandemic Influenza A (H1N1) 2009 Vaccines: A Multinational Self-Controlled Case Series in Europe. PLoS ONE. 3 Jan 2014;9(1):e82222.
27. Avillach P, Dufour J-C, Diallo G, Salvo F, Joubert M, Thiessard F, et al. Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EU-ADR project. J Am Med Inform Assoc. May2013;20(3):446–52.
28. Bauer-Mehren A, van Mullingen EM, Avillach P, Carrascosa MDC, Garcia-Serna R, Piñero J, et al. Automatic filtering and substantiation of drug safety signals. PLoS Comput Biol.2012;8(4):e1002457.
29. http://www.imi.europa.eu/content/emif. Accessed September 2015
30. Mini-Senetinel. www.mini-sentinel.org. Accessed September 2015
31. Raebel MA, Penfold R, McMahon AW, Reichman M, Shetterly S, Goodrich G, et al. Adherence to guidelines for glucose assessment in starting second-generation antipsychotics. Pediatrics. 2014;134(5):e1308–14.
32. Carnahan RM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned. Pharmacoepidemiology and Drug Safety. Jan 2012;21:82–9. 33. Cutrona SL, Toh S, Iyer A, Foy S, Cavagnaro E, Forrow S, et al. Design for validation of acute myocardial infarction cases in Mini-Sentinel. Pharmacoepidemiology and Drug Safety. Jan 2012;21:274–81.
34. Cutrona SL, Toh S, Iyer A, Foy S, Daniel GW, Nair VP, et al. Validation of acute myocardial infarction in the Food and Drug Administration's Mini-Sentinel program. Pharmacoepidemiol Drug Saf. Jan 2013;22(1):40–54.
35. McClure DL, Raebel MA, Yih WK, Shoaibi A, Mullersman JE, Anderson-Smits C, et al. Mini-Sentinel methods: framework for assessment of positive results from signal refinement: FRAMEWORK FOR ASSESSMENT OF POSITIVE RESULTS FROM SIGNAL REFINEMENT. Pharmacoepidemiology and Drug Safety. 2014;23(1):3–8.
36. Raebel MA, Haynes K, Woodworth TS, Saylor G, Cavagnaro E, Coughlin KO, et al. Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel: THE MINI-SENTINEL LABORATORY RESULTS TABLE. Pharmacoepidemiology and Drug Safety. 2014;23(6):609–18.
37. Rassen JA, Schneeweiss S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. Pharmacoepidemiology and Drug Safety. Jan2012;21:41–9.
38. Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership, Stat Med. 2012 Dec 30;31(30):4401-15.
39. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012 Jan- Feb;19(1):54-60.

40. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inform. Aug2012;45(4):689–96.

41. Hartzema AG, Reich CG, Ryan PB, Stang PE, Madigan D, Welebob E, et al. Managing Data Quality for a Drug Safety Surveillance System. Drug Safety. 29 Oct2013;36(S1):49–58.

42. Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, et al. A Comparison of the Empirical Performance of Methods for a Risk Identification System. Drug Safety. 29 Oct 2013;36(S1):143–58.

43. Observational Health Data Sciences and Informatics. ohdsi.org. Accessed September 2015

44. Buja A, Gini R, Visca M, Damiani G, Federico B, Francesconi P, et al. Prevalence of chronic diseases by immigrant status and disparities in chronic disease management in immigrants: a population-based cohort study, Valore Project. BMC Public Health. 24 may 2013;13(1):504.

45. Buja A, Damiani G, Gini R, Visca M, Federico B, Donato D, et al. Systematic Age-Related Differences in Chronic Disease Management in a Population-Based Cohort Study: A New Paradigm of Primary Care Is Required. PLoS ONE. 14 Mar 2014;9(3):e91340.

46. Buja A, Gini R, Visca M, Damiani G, Federico B, Donato D, et al. Need and disparities in primary care management of patients with diabetes. BMC Endocrine Disorders. 10 Jul 2014;14(1):56.

47. Visca M, Donatini A, Gini R, Federico B, Damiani G, Francesconi P, et al. Group versus single handed primary care: a performance evaluation of the care delivered to chronic patients by Italian GPs. Health Policy. Nov 2013;113(1-2):188–98.

48. Gini R, Francesconi P, Mazzaglia G, Cricelli I, Pasqua A, Gallina P, et al. Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey. BMC Public Health. 9 Jan2013;13(1):15.

49. Gini R, Schuemie MJ, Lapi F, Cricelli I, Pasqua A, et al. Can Italian healthcare administrative databases be used to compare regions with respect to compliance with standards of care for chronic diseases? PLoS ONE. May 2014;9(5):e95419.

50. Cali A, Calvanese D, De Giacomo G, Lenzerini M. Data integration under integrity constraints. Information Systems, 29(2):147–163, 2004.

51. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation. PLoS ONE. November 2014;9(11):e110900.

52. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care.2013 Aug;51(8 Suppl 3):S22-9.

53. thematrix.isti.cnr.it. Accessed September 2015.

54. Avillach P, Coloma PM, Gini R, Schuemie M, Mougin F, Dufour J-C, et al. Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. J Am Med Inform Assoc. 6 Sep2012

55. Cazzola W. Domain-Specific Languages in Few Steps: The Neverlang Approach. In Proceedings of the 11th International Conference on Software Composition (SC'12), Prague, Czech Republic, May-June2012, LNCS 7306, pp. 162–177, Springer.

56. Cazzola W, Vacchi E. Neverlang 2: Componentised Language Development for the JVM. In Proceedings of the 12th International Conference on Software Composition (SC'13), Budapest, Hungary, June 2013, LNCS 8088, pp. 17–32, Springer.

57. Adler-Milstein J, Ronchi E, Cohen GR, Winn LAP, Jha AK. Benchmarking health IT among OECD countries: better data for better policy. J Am Med Inform Assoc. 1 Jan 2014;21(1):111–6. 58. Schuemie MJ, Gini R, Coloma PM, Straatman H, Herings RMC, Pedersen L, et al. Replication of the OMOP Experiment in Europe: Evaluating Methods for Risk Identification in Electronic Health Record Databases. Drug Safety. Oct 2013;36(S1):159–69.

59. Toh SS, Gagne JJP, Rassen JAS, Fireman BH, Kulldorff M, Brown JS. Confounding Adjustment in Comparative Effectiveness Research Conducted Within Distributed Research Networks. Medical Care. 2013.

60. Toh S, Reichman ME, Houstoun M, Ding X, Fireman BH, Gravel E, et al. Multivariable confounding adjustment in distributed data networks without sharing of patient-level data. Pharmacoepidemiol Drug Saf. 2013;22(11):1171–7.

61. Toh SS, Shetterly SM, Powers JDM, Arterburn D. Privacypreserving Analytic Methods for Multisite Comparative Effectiveness and Patient-centered Outcomes Research. Medical Care. 2014. 2014;52(7):664–8.

62. Hernán MA. With great data comes great responsibility: publishing comparative effectiveness research in EPIDEMIOLOGY. Epidemiology. May2011;22(3):290–1.

63. Hansen RA, Gray MD, Fox BI, Hollingsworth JC, Gao J, Zeng P. How Well Do Various Health Outcome Definitions Identify Appropriate Cases in Observational Studies? Drug Safety. 29 Oct2013;36(S1):27–32.

64. Valkhoff VE, Coloma PM, Lapi F, Gini R, Nielsson MS, Mosseveld M, Molokhia M, Schuemie MJ, Sturkenboom MCJM, Trifirò G. Positive predictive value for upper gastrointestinal bleeding in four health care databases using different coding systems in the EU-ADR project. Presented at the Digestive Disease Week, San Diego, California, May 19–22, 2012.

65. Reich CG, Ryan PB, Schuemie MJ. Alternative Outcome Definitions and Their Effect on the Performance of Methods for Observational Outcome Studies. Drug Safety. Oct 2013;36(S1):181–93.

66. Maro JC, Brown JS, Dal Pan GJ, Kulldorff M. Minimizing signal detection time in postmarket sequential analysis: balancing positive predictive value and sensitivity. Pharmacoepidemiol Drug Saf. 2014

67. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. Statistics in Medicine. Jul 2013

68. Popper K. Conjectures and refutations: the growth of scientific knowledge. Routledge 1963. London. 21

CHAPTER 7

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

Giuseppe Roberto[1], Ingrid Leal[2], Naveed Sattar[3], A. Katrina Loomis[4], Paul Avillach[2,5], Peter Egger[6], Rients van Wijngaarden[7], David Ansell[8], Sulev Reisberg[9], Mari-Liis Tammesoo[10, 11], Helene Alavere[10, 11], Alessandro Pasqua[12], Lars Pedersen[13], James Cunningham[14], Lara Tramontan[15], Miguel A. Mayer[16], Ron Herings[7], Preciosa Coloma[2], Francesco Lapi[1], Miriam Sturkenboom[2], Johan van der Lei[2], Martijn J. Schuemie[17, 18], Peter Rijnbeek[2] and Rosa Gini[1]

[1]Regional Agency for Healthcare Services of Tuscany, Epidemiology unit, Florence, Italy[2] Department 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]Department of Biomedical Informatics, Harvard Medical School & Children's Hospital Informatics Program, Boston Children's Hospital, Boston, Massachusetts
[6]GlaxoSmithKline, Worldwide Epidemiology GSK, Stockley Park West, Uxbridge, United Kingdom
[7]PHARMO Institute for Drug Outcomes Research, Utrecht, Netherlands
[8]The Health Improvement Network, Cegedim Strategic Data Medical Research Ltd, London, United Kingdom
[9]Quretec, Software Technology and Applications Competence Center, University of Tartu, Tartu, Estonia
[10]Estonian Genome Center, University of Tartu, Tartu, Estonia
[11]Tartu University Hospital, Tartu, Estonia
[12]Health Search, Italian College of General Practitioners and Primary Care, Firenze, Italy
[13]Department of Clinical Epidemiology, Aarhus University Hosptial, Aarhus, Denmark
[14]University of Manchester, Manchester, United Kingdom
[15]Arsenàl.IT Consortium, Veneto's Research Centre for eHealth Innovation, Treviso, Italy
[16]Hospital del Mar Medical Research Institute (IMIM) and Universitat Pompeu Fabra, Barcelona, Spain
[17]Janssen Research & Development, Epidemiology, Titusville, New Jersey, United States
[18]Observational Health Data Sciences and Informatics, New York, New York, United States.

Roberto G, Leal I, Sattar N, Loomis AK, Avillach P, Egger P, van Wijngaarden R, Ansell D, Reisberg S, Tammesoo ML, Alavere H, Pasqua A, Pedersen L, Cunningham J, Tramontan L, Mayer MA, Herings R, Coloma P, Lapi F, Sturkenboom2ì M, van der Lei J, Schuemie MJ, Rijnbeek P, Gini R. Identifying cases of type 2 diabetes in heterogeneous data sources: strategy from the EMIF project. PLoS ONE 2016;11(8):e0160648.

ABSTRACT

Due to the heterogeneity of existing European sources of observational healthcare data, data source-tailored choices are needed to execute multi-data source, multi-national epidemiological studies. This makes transparent documentation paramount.

In this proof-of-concept study, a novel standard data derivation procedure was tested in a set of heterogeneous data sources. Identification of subjects with type 2 diabetes (T2DM) was the test case.

We included three primary care data sources (PCDs), three record linkage of administrative and/or registry data sources (RLDs), one hospital and one biobank. Overall, data from 12 million subjects from six European countries were extracted. Based on a shared event definition, sixteeen standard algorithms (*components*) useful to identify T2DM cases were generated through a top-down/bottom-up iterative approach. Each component was based on one single data domain among diagnoses, drugs, diagnostic test utilization and laboratory results. Diagnoses-based components were subclassified considering the healthcare setting (primary, secondary, inpatient care). The Unified Medical Language System was used for semantic harmonization within data domains.

Individual components were extracted and proportion of population identified was compared across data sources. Drug-based components performed similarly in RLDs and PCDs, unlike diagnoses-based components. Using components as building blocks, logical combinations with AND, OR, AND NOT were tested and local experts recommended their preferred data source-tailored combination. The population identified per data sources by resulting algorithms varied from 3.5% to 15.7%, however age-specific results were fairly comparable. The impact of individual components was assessed: diagnoses-based components identified the majority of cases in PCDs (93-100%), while drug-based components were the main contributors in RLDs (81-100%).

The proposed data derivation procedure allowed the generation of data source-tailored case-finding algorithms in a standardized fashion, facilitated transparent documentation of the process and benchmarking of data sources, and provided bases for interpretation of possible inter-data source inconsistency of findings in future studies.

INTRODUCTION

In recent years, an increasing number of projects have been focusing on re-using existing electronic health records (EHR) for clinical research.[1] In particular, huge efforts have been made to combine health data from isolated environments and perform valid multi-data source observational studies.[2, 3]
In this context, the European Medical Information Framework (EMIF) project was launched with the main objective of building an infrastructure for the efficient re-use of existing European health care data for epidemiological research (http://www.emif.eu/). Within the project, a federation of heterogeneous sources of real world data (e.g. administrative, hospital or primary care databases, disease registries, biobanks), currently collecting health information on around 52 million European citizens, collaborate in the EMIF-Platform whose focus is the consistent exploitation of currently available patient-level data to support novel research. One of the main challenges for the EMIF-Platform is to deal with the heterogeneous characteristics of the participating data sources and facilitate the execution of high quality multi-national, multi-data source observational studies based on populations with otherwise unconceivable sample sizes and follow-up time span.
In general, different strategies can be adopted to identify a population of interest from a single source of EHR.[4, 5] The choice of a particular case-finding algorithm is generally driven by both the specific research question and the data source peculiarities.[6] The chosen algorithm, however, can significantly affect the characteristics of the cases identified [4, 5] and, for this reason, should be carefully taken into account when discussing study results.
In multi-data source studies, tailored choices may be necessary [6-8], and the diversity of local case-identification algorithms may increase along with the heterogeneity of the data sources involved [6, 9, 10]. A transparent process of documentation and evaluation of local case-finding algorithms becomes paramount for the correct interpretation of study results as well as for the discussion of possible inter-data source inconsistency of study findings [10-12].
It must be noted that data sources available to study European populations are much more heterogeneous than data sources from a single country, such as the United States [10]. Therefore, in order to address this issue, the EMIF-Platform designed a novel standard procedure for data derivation which leverages the experience gained from previous European multi-national, multi-data source

studies [2, 3, 9, 13]. In this proof-of-concept study, the identification of type 2 diabetes mellitus (T2DM), a common chronic condition with important implications for future health[14], was used as a test case.

MATERIALS AND METHODS

Data sources

Eight European data sources collecting health care information on around 20 million subjects from six different countries participated to this study (Table 1). Three were primary care data sources (PCDs), three were record linkage systems of different registries (RLDs), one was a hospital data source (HD) and one was a biobank (BD). In specific, the three primary care data sources were the Health Search IMS Health LPD database (HSD, Italy),[9, 15] the Integrated Primary Care Information database (IPCI, The Netherlands)[16] and The Health Improvement Network database (THIN, UK), in which the general practitioners (GPs) function as data keeper of all patient's medical information.[17] The three record linkage data sources were the Aarhus University Hospital (AUH, Aarhus, Denmark),[18, 19] PHARMO (PHARMO, The Netherlands)[20] and the Regional Health Authority of Tuscany (ARS, Italy),[9, 15] which collect data from different sources (e.g. hospital discharge records, death registries, drug dispensing and procedures). The HD was the Information System of Parc de Salut Mar Barcelona (IMASIS, Spain) that records information from routine healthcare activities of Hospital del Mar of Barcelona.[21, 22] The BD was the Estonian Genome Center of University of Tartu (EGCUT, Estonia) in which information from interviews of voluntary donors of biological samples is collected through standard questionnaires.[23] EGCUT is the only cross-sectional data source included in this study. In all data sources except the Spanish HD, IMASIS, information on a representative sample of the population living in the corresponding geographic area are collected. In the Italian PCD and in the Estonian BD only adult population is represented (>14 and >18 years of age, respectively). The information in the corresponding databases is recorded using different coding systems. Diagnoses are coded according to the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) or ICD-10 (10th version), the International Classification of Primary Care (ICPC), READ or are as free text. Prescriptions/dispensings are

Table 1. Data sources' characteristics*

Data source (Original organization's acronym)	**Type of data source**	**Catchment area**	**Cumulative number of participants in the database**	**Average follow-up time**	**Diagnoses** (setting, coding system)	**Medication** (coding system)	**Diagnostic procedures/ tests** (coding system)	**Laboratory results** (coding system for measurments)
RLD-I (ARS)	Record linkage system	Tuscany (Italy)	5 millions	9 years	Inpatient, ICD9CM	ATC	ICD9CM or local terminology	-
RLD-DK (AUH)	Record linkage system	The northern and central region of Jutland. (Denmark)	2.3 millions	13 years	Inpatient, secondary care ICD10	ATC	NOMESCO	-
RLD-N (PHARMO)	Record linkage system	Netherlands (Certain regions, mainly South East and North-West)	10 millions	10 years	Inpatient, ICD9CM	ATC	Local terminology	Local terminology
PCD-I (HSD)	Primary care	Italy	2.3 millions	10 years	Primary care, ICD9CM	ATC	Local terminology	Local terminology
PCD-UK (THIN)	Primary care	United Kingdom	12 millions	9 years	Primary care, RCD	ATC	Local terminology	Local terminology
PCD-N (IPCI)	Primary care	Netherlands	2.8 millions	3 years	Primary care, ICPC/free text	ATC	Local terminology	Local terminology
HD (IMASIS)	Hospital	Barcelona (three city districts)	1.5 millions	5 years	Admissions, outpatients, major ambulatory surgery and emergency room ICD9CM	Local terminology & the Spanish Medicines Agency codes	ICD9CM	Local terminology
BD (EGCUT)	Biobank	Estonia	52000	Not applicable	Primary care/Self reported, ICD10	ATC	Local terminology	Local terminology

*Information reported in the table is updated at January 2013.

coded according to the Anatomical Therapeutic Chemical classification (ATC) or BNF/Multilex. The majority of the data sources collect records concerning the utilization of diagnostic procedures and laboratory results. The coding of these data domains are based on local service terminologies.

Study population and design

In each participating data source the study population corresponded to all active subjects on January the 1st 2012 (reference date) that at the same date had ≥16 years of age. Due to sample size issues, exception was made for EGCUT in which January the 1st 2009 was considered as the reference date.

A descriptive, cross-sectional, retrospective multi-database study was performed. Patients with T2DM were identified within the populations selected from the participating data sources by using different case-finding algorithms.

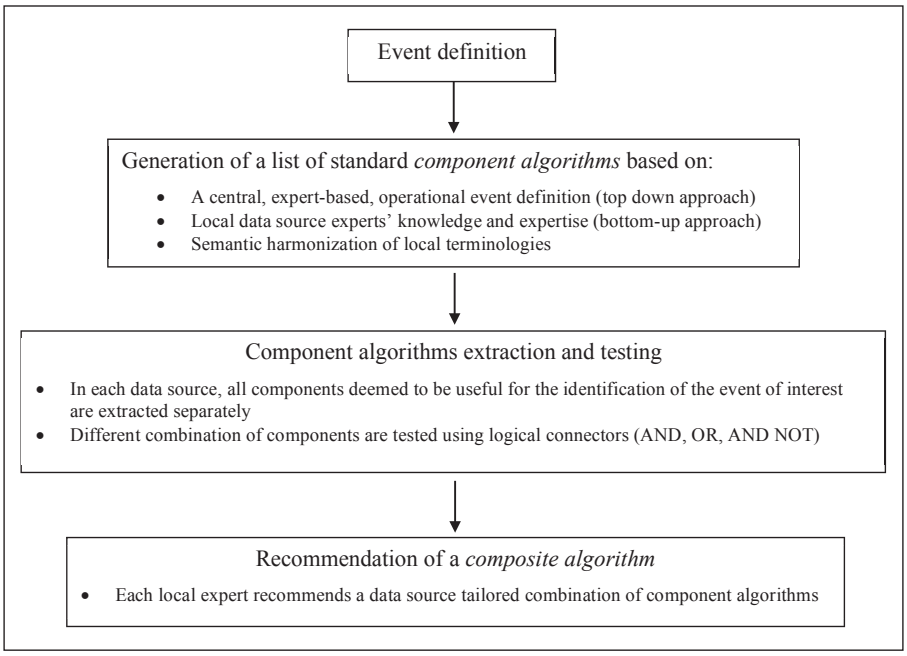
Event definition and generation of a list of component algorithms

T2DM is a chronic clinical condition characterized by hyperglycemia due to insulin resistance and a progressive deficiency in insulin production.[24] It represents the most common form of diabetes, comprising about 90% of all cases of diabetes worldwide.[25] Diagnosis and follow-up of T2DM is based on laboratory tests for blood glucose measurements and treatment includes life style interventions (i.e. diet and physical exercise) and use of medications.[26]

To identify subjects with T2DM in a healthcare data source, information from one or more data domains may be available. Diagnoses and/or records collecting information on routine patients' clinical care and follow-up, such as drug prescriptions, utilization of diagnostic tests and laboratory results, can be used,[4, 27, 28] so that combining data from one or more of these domains, different case-finding strategies, with different sensitivity and positive predictive value (PPV), can be obtained.[4]

As the first step of the data derivation procedure, a shared clinical definition of T2DM was adopted (Figure 1) and defined according to the ESC/EASD guideline.[29]

Figure 1. The standard procedure for data derivation



Subsequently, a list of standard algorithms useful to identify cases of T2DM in the selected data sources was generated. Standard algorithms, referred to as "component algorithms", were defined as rules to identify subjects with a defined pattern of records selected from a single data domain. For the identification of T2DM, a total of four data domains were concerned: diagnoses (DIAG), drug prescription/dispensing (DRUG), utilization of a diagnostic test (TEST) or laboratory results (LABVAL). Component algorithm could be intended as inclusion, exclusion or refinement criteria. Two sources of knowledge were leveraged and integrated for the design of component algorithms: a central expert-based clinical and operational definition of T2DM (top-down engineering) and the expertise provided by local data source experts with respect to the identification of T2DM cases in their own data source (bottom-up learning). [8] As already described in greater details in a previous published paper,[3] the Unified Medical Language System (UMLS) was used to build a shared semantic foundation across the different coding systems: medical concepts pertinent to the clinical and operational definition of T2DM were identified and projected to local terminologies. The final list of local codes, strings and free text keywords

was obtained through an iterative process involving local experts' feedback. Each component algorithm was fully described by two additional rules: the first was the pattern of records that triggers identification of the event (for instance: at least two records in the same calendar year), and the second concerned the criteria to identify the case's index date (e.g. date of the first record).

Data extraction and analysis: "the component algorithm strategy"

A distributed network approach was adopted in EMIF to allow partners for maintaining control of their data and to benefit from local data source experts consultation on the appropriate use of data and interpretation of results.[2, 3] Local experts were asked to select and extract all component algorithms considered useful to identify T2DM cases in their data source. All person-time available up to the reference date was considered for algorithm application. Extracted data were prepared to be inputted in Jerboa, a custom-built software developed in the EU-ADR project[2] which was run locally to standardize the data aggregation process. After providing formal approval, local data source experts uploaded aggregated analytical datasets to a common virtual machine.

Using a custom-built analysis tool (a Microsoft Access interface for Stata [StataCorp. 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP] and LaTeX [https://www.latex-project.org/]), local experts could test the extracted components in any possible logical combinations by using Boolean operators (i.e. AND, OR, AND NOT). This strategy, we referred to as *"the component algorithm strategy",* allowed local experts to build more complex case-finding strategies (*composite algorithms)* by combining two or more of the extracted components as a mean of inclusion, exclusion or refinement criteria. Testing different combination of components, local experts could choose a particular composite algorithm that they recommended for the identification of T2DM in their data source. A comment describing the reasons behind the choice was recorded together with an estimate, either objective or subjective, of the expected sensitivity and PPV. This information was stored and intended as a source of reusable knowledge.

Presentation of results

Results from the application of individual components and recommended composite algorithms were compared across data sources and presented as age-specific percentages of subjects identified in the study population of the corresponding data source.

In each participating data source, the impact of extracted component was assessed with respect to the total number of subjects identified using the recommended composite algorithm, which was considered as the *reference case population*. For this purpose, we calculated: i) the percentage of subjects identified by each component in the reference case population and ii) the prevalence rate ratio (PRR) of subjects identified by the recommended composite algorithm with and without the use of the tested component as additional inclusion criteria, i.e. PRR=((N in tested component OR in recommended composite algorithm)/ N in recommended composite algorithm)-1

No identifiable human data were shared for this study. Permission for both re-use of the data analyzed in this study as well as for publication of the results obtained was granted by each participating organizations' review board.

The full protocol of the research project is publicly available on the electronic register of observational studies of the European Network of Centers for Pharmacoepidemiology and Pharmacovigilance (http://www.encepp.eu/encepp/viewResource.htm?id=11158)

RESULTS

Since this was a proof-of-concept study, results presented here are not intended as estimates of disease frequency.

Overall, the EMIF-Platform provided for this study aggregated health data from around 12 million European citizens.

The size of the study populations selected from the participating data sources ranged from 1600 to 3.4 million subjects. Components algorithms included in at least one recommended composite algorithms are reported in Table 2.

In Figure 2 four examples of comparisons of age band-specific results from individual component algorithms across data sources are shown. The full list of comparisons concerning all those components extracted from at least two data sources and included in at least one recommended composite algorithm are available as supporting in formation in Figure 1S. As for DIAG-based components very different performances were associated to the healthcare setting of data collection (primary, secondary, inpatient care). The component DRUG_ORAL (i.e. ≥2 records of non-insulin antidiabetic drugs utilization in

Table 2. Component algorithms description

Component algorithm acronym	Algorithm description	Record retrieval rules*	Case's index date
DIAG_T2DM_PC	Patients who have ≥1 diagnoses of T2DM recorded in a primary care setting	Records of (Diabetes type 2) occurs in [diagnosis fields] of [tables collected during primary care]	1st record
DIAG_T2DM_SC	Patients who have ≥1 diagnoses recorded in a secondary care setting	Records of (Diabetes type 2) occurs in [diagnosis fields] of [tables collected during secondary care]	1st record
DIAG_T2DM_INP	Patients who ≥1 diagnoses recorded during a hospital admission	Records of (Diabetes type 2) occurs in [diagnosis fields] of [tables collected during inpatient care]	1st record
DIAG_DMUNSPEC	Patients who ≥1 diagnoses of unspecified diabetes recorded in primary, secondary, or inpatients care	Records of (Diabetes unspecified) occurs in [diagnosis fields] of [tables collected in primary, secondary, or inpatients care]	1st record
DIAG_DMUNSPEC_OTH	Patients who have ≥1 diagnoses recorded in a setting other than primary, secondary, or inpatients care	Records of (Unspecified diabetes) occurs in [diagnosis fields] of [tables collected in other settings]	1st record
DIAG_T1DM	Patients who have ≥1 diagnoses of T1DM recorded in any care setting	Records of (Diabetes mellitus type 1) occurs in [diagnosis fields] of [any table collecting diagnoses]	1st record
DIAG_EXCL	Patients who have ≥1 diagnoses of conditions excluding T2DM other than T1DM recorded in any care setting	Records of ((Metabolic problems around pregnancy) OR (Metabolic/pancreatic problems, non type 2 diabetes) OR (Polycystic Ovary Syndrome) occurs in [diagnosis fields] of [any table collecting diagnoses]	1st record
DRUG_INSULIN_ONE	Patients who have ≥1 recorded prescriptions/ dispensings of insulin	Records of (Insulins and analogues) occurs in [ATC field] of [drugs tables]	1st record
DRUG_INSULIN	Patients who have ≥2 recorded prescriptions/ dispensings of insulin in a calendar year	Records of (Insulins and analogues) occurs in [ATC field] of [drugs tables]	2nd record
DRUG_ORAL_ONE	Patients who have ≥1 recorded prescriptions/ dispensings of non-insulin antidiabetic drugs	Records of (Drugs used in diabetes, excl insulin) occurs in [ATC field] of [drugs tables]	1st record
DRUG_ORAL	Patients who ≥2 prescriptions/dispensings of non-insulin antidiabetics in a calendar year	Records of (Drugs used in diabetes, excl insulin) occurs in [ATC field] of [drugs tables]	2nd record
TEST_GLUCO5_1YR	Patients who have ≥5 records of utilization of blood glucose measurements within 1 year	Records of (Blood glucose measurement) occurs in [code of test field] of [tables collecting laboratory test results or dispensings]	5th record

Table 2. Continued

Component algorithm acronym	Algorithm description	Record retrieval rules*	Case's index date
TEST_ GLUCO2_ PYEAR_5YRS	Patients who have ≥2 records of utilization of blood glucose measurements per year for 5 consecutive years	Records of (Blood glucose measurement) occurs in [code of test field] of [tables collecting laboratory test results or dispensings]	2nd record
LABVAL_ HbA1c	Patients who have ≥2 laboratory results recorded from a glycated hemoglobin test higher than 6.5% (48 mmol/mol)	Records of (Glycated Haemoglobin) occurs in [code of test field] of [tables collecting laboratory test results] AND [result field] of the same record is higher than 6.5% (or 48 mmol/mol, according to unit of measurement adopted in the table)	2nd record
LABVAL_ FAST_GLUC	Patients who have ≥2 laboratory results recorded from a fasting plasma glucose measurement higher than 126 mg/dl)	Records of (Fast gluc) occurs in [code of test field] of [tables collecting laboratory test results] AND [result field] of the same record is higher than 126 mg/dl	2nd record
LABVAL_ LCURVE_ GLUC	Patients who have ≥2 laboratory results recorded from a glucose tolerance test higher than 200 mg/dl	Records of (LcurveGLuc) occurs in [code of test field] of [tables collecting laboratory test results] AND [result field] of the same record is higher than 200 mg/dl	2nd record

*Codes and free text keywords corresponding to the medical concepts embedded in component algorithms (in brackets) are reported in Supplemental Table 1.

Figure 2. Comparison of results from individual component algorithms: four examples

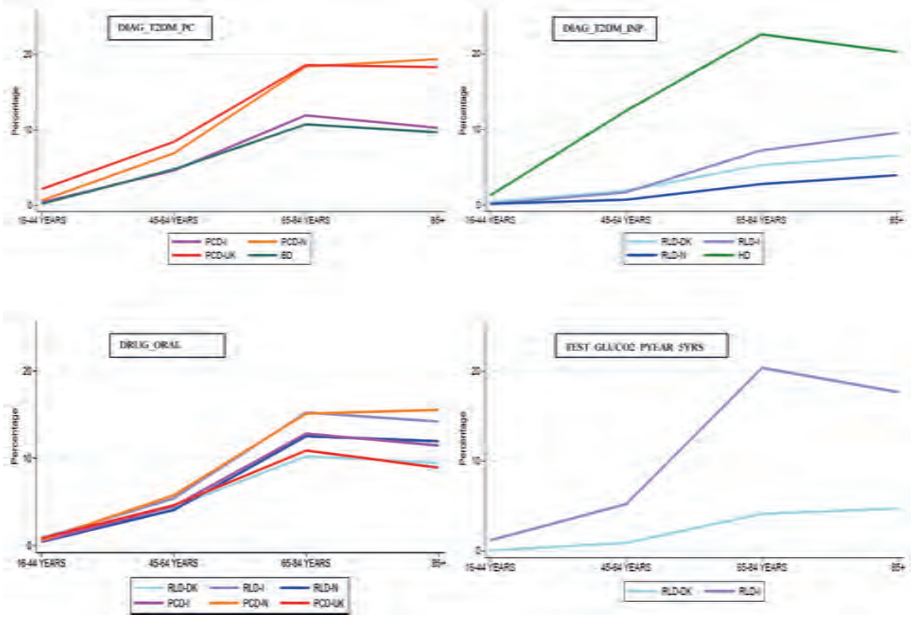
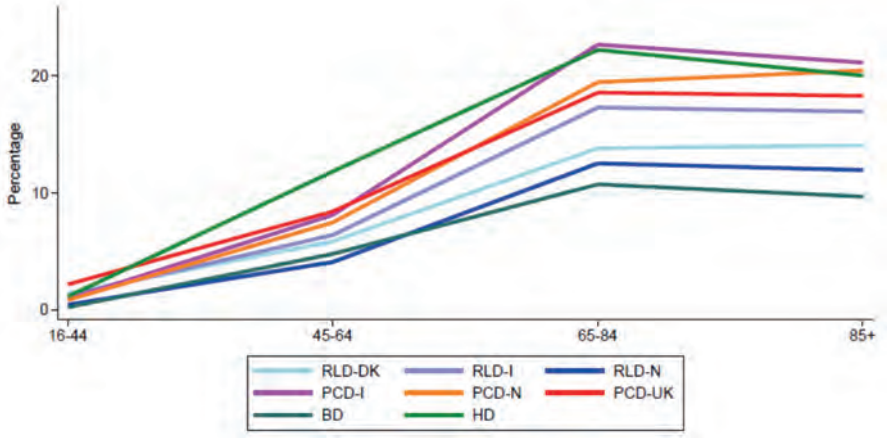


Figure 3. Recommended composite algorithms: age band-specific percentages of subjects identified on the relevant total study population



Data Source	Recommended composite algorithm	Comment of the local expert	Sensitivity	PPV
PCD-I	(DIAG_ DMUNSPEC **OR** LABVAL_ HbA1c **OR** LABVAL_ FAST_GLUC **OR** LABVAL_ LCURVE_ GLUC) **AND NOT** (DIAG_T1DM))	The chosen composite algorithm was validated in HSD in a study that is undergoing publication and found very high PPV (around 100% in the validation sample). Due to the nature of the data source and the very broad algorithm, sensitivity must be very high as well.	≥.9	≥.9
PCD-N	DIAG_T2DM_ PC **OR** DRUG_ ORAL_ONE	T2DM is identified via diagnosis codes (DIAG_PC) or utilization of specific drugs for T2DM (non-insulin antidiabetic drugs). It has been observed that there is no substantial difference between selection of patients with one (DRUG_ORAL_ONE) or at least two drug prescriptions (DRUG_ORAL). Some subjects with T2DM diagnosis also have a record for type 1 diabetes. Some GPs record type 1 diabetes to indicate insulin dependence (regardless of diabetes type) due to the former type 1 diabetes name 'insulin-dependent diabetes'.	≥.9	≥.9
PCD-UK	DIAG_T2DM_ PC	Adding DRUG-based, as well as other possible inclusion criteria, components does not add almost anybody.	≥.9	≥.9

Figure 3. Continued

Data Source	Recommended composite algorithm	Comment of the local expert	Sensitivity	PPV
RLD-DK	(DIAG_ T2DM _INP **OR** DIAG_T2DM_ SC) **OR** ((TEST_ GLUCO5_1YR **OR** DRUG_ORAL **OR** DRUG_ INSULIN **OR** TEST_ GLUCO2 PYEAR 5YRS) **AND NOT** (DIAG_T1DM **OR** DIAG_EXCL))	The chosen strategy to identify T2DM cases is similar to the strategy which is regularly used in this data source to identify cases of unspecified diabetes. This is the strategy of the Danish National Diabetes registry, which has been repeatedly validated. A recent study published in 2015 estimated that sensitivity and PPV are 95 and 80%, respectively. When adapting this algorithm to the case of T2DM we decided to change some elements of the validated strategy. The main differences are: we used type 1 diabetes diagnoses as exclusion criteria, we used diagnoses of T2DM rather than diagnoses of unspecified diabetes, we avoided chiropody for diabetics as inclusion criteria and we did not exclude cases of gestational diabetes. As for insulin and other antidiabetic drugs, we used two prescription in one year as inclusion criteria rather than two prescription recorded at any time. The expected sensitivity and PPV of the chosen algorithm are possibly slightly lower but still very close to those of the validated algorithm.	≥.9	>.7 and <.9
RLD-I	DIAG_INP_ T2DM **OR** DRUG_ORAL **OR** DRUG_ INSULIN **OR** DIAG_T2DM_ OTH	From a validation study, the sensitivity of this algorithm is 76% and PPV 86%, excluding subjects with a record of type 1 diabetes does not improve the extraction validity. DIAG_ T2DM_OTH refers to an Italian table: disease-specific exemptions from copayment to healthcare.	>.7 and <.9	>.7 and <.9
RLD-N	DRUG_ORAL	The chosen strategy to identify T2DM patients has a very high PPV (95%). The reason we have decided to use only this method and not include other components we extracted is that the identification of T2DM patients based on the use of oral antidiabetics is extensively tested and validated within GPs. We are aware that this approach is limited to identify treated patients only, but this is deliberate. Although a patient could be classified as a T2DM patient given the strict clinical definition that was chosen in this process (i.e. HbA1c above a certain threshold), the fact that the patient is not treated begs the question whether it should be classified as a prevalent patient. The sensitivity of this strategy with respect to the wider class of patients matching the clinical definition is between 70 and 90%.	>.7 and <.9	≥.9
BD	DIAG_T2DM_ PC	Using any other extracted component algorithm as additional inclusion criteria do not add any patients if those with a concomitant diagnosis of type 1 diabetes are excluded.	≥.9	≥.9
HD	DIAG_T2DM_ INP **AND NOT** DIAG_T1DM	The chosen algorithm uses as inclusion criterion inpatients diagnoses of T2DM and exclude patients with diagnosis of type 1 diabetes. Other extracted components did not significantly affect the size of the population of cases identified.	≥.9	≥.9

one calendar year) and DRUG_INSULIN (i.e. ≥2 records of insulin utilization in one calendar year) were extracted in all the participating PCDs and RLDs and resulted in a comparable age band-specific percentage of subjects identified in the respective study populations.

The data source tailored recommended composite algorithms are shown in Figure 3 together with the comments of the local experts. The PCD from UK and the BD from Estonia adopted the component algorithm based on T2DM diagnoses from primary care (DIAG_T2DM_PC) only as the recommended choice. The HD from Spain excluded from the pool of subjects identified through inpatients diagnoses of T2DM (DIAG_T2DM_INP) those with a recorded diagnosis of type 1 diabetes (DIAG_T1DM). Three data sources (the PCD from Italy and the RLDs from Denmark and Italy) adopted
complex composite algorithms, based on the results of previous validation studies [27, 30, 31]. The Dutch PCD added a sensitive pattern of utilization of non-insulin antidiabetic drugs (i.e. DRUG_ORAL_ONE) as inclusion criterion, due to the observed low sensitivity of the DIAG-based algorithm DIAG_T2DM_PC in this data source. The Dutch RLD chose to include only subjects utilizing non-insulin antidiabetics, because the available DIAG-based component that used diagnoses from inpatients setting was considered unreliable by local experts.

Through the application of the recommended composite algorithm, the lowest percentage of study population was identified in the Estonian BD, 3.5%, while the highest in the Spanish HD, 15.7%. In the RLDs it ranged from 4.1% to 7.5% while in PCDs from 6.8% to 8.6%. The age band-specific percentages of the total case populations identified using the recommended composite algorithms showed more comparable results across all participating data sources (Figure 3). The expected sensitivity of the recommended composite algorithms, as reported by local experts either form previous validation studies or from subjective judgement, was >0.9 in all data sources except for the Italian and Dutch RLDs for which a sensitivity between 0.7 and 0.9 was expected. As for PPV, the Italian and Danish RLDs only reported an expected value ranging from 0.7 and 0.9 while for the remaining data sources the figure was >0.9.
The union of any extracted DIAG-based component among the five intended as inclusion criteria identified from 93 to 100% of the reference case population in PCDs (Table 3), 100% in both BD and HD, and from 15% to 73% in RLDs.

Table 3. Impact of extracted component algorithms on total case population identified in each participating data source through the application of the relevant recommended composite algorithm.

COMPONENT ALGORITHMS (B)°		RECOMMENDED COMPOSITE ALGORITHMS (A)							
		RLD-I	RLD-DK	RLD-N	PCD-UK	PCD-N	PCD-I	BD	HD
	N	3391177	1372883	1405220	3278013	992924	945691	22430	15713
	N in A	254045	77616	57712	253197	67096	81658	779	2466
	% of A in N	7.5	5.7	4.1	7.7	6.8	8.6	3.5	15.7
DIAG_T2DM_PC ≥1 diagnosis from primary care	N in B	n.e.	n.e.	n.e.	253197	62191	43438	779	n.e.
	% of B in A	-	-	-	100.0%	92.7%	52.6%	100.0%	-
	PRR if B added	-	-	-	+0.0%	+0.0%	+0.6%	+0.0%	-
DIAG_T2DM_INP ≥1 T2DM diagnosis from inpatient care	N in B	95303	27887	13098	n.e.	n.e.	n.e.	n.e.	2520
	% of B in A	37.5%	35.9%	15.1%	-	-	-	-	100.0%
	PRR if B added	+0.0%	+0.0%	+7.6%	-	-	-	-	+2.2%
DIAG_T2DM_SC ≥1 T2DM diagnosis from secondary care	N in B	n.e.	35744	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
	% of B in A	-	46.1%	-	-	-	-	-	-
	PRR if B added	-	+0.0%	-	-	-	-	-	-
DIAG_DMUNSPEC ≥1 unspecified diabetes diagnosis from any healthcare setting	N in B	191999	n.e.	n.e.	n.e.	n.e.	79035	n.e.	n.e.
	% of B in A	73.2%	-	-	-	-	94.3%	-	-
	PRR if B added	+2.4%	-	-	-	-	+2.5%	-	-
DIAG_DMUNSPEC_OTH ≥1 unspecified diabetes diagnosis from co-payment exemption	N in B	149806	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.	n.e.
	% of B in A	59.0%	-	-	-	-	-	-	-
	PRR if B added	+0.0%	-	-	-	-	-	-	-
DIAG_T1DM ≥1 type 1 diabetes diagnoses from any healthcare setting	N in B	18147	17896	n.e.	n.e.	8816	2050	164	78
	% of B in A	6.9%	18.1%	-	-	8.8%	0.0%	2.8%	0.0%
	PRR if B added	+0.2%	+4.9%	-	-	+4.3%	+2.5%	+18.2%	+3.2%

Table 3. Continued

DIAG_EXCL ≥1 diagnoses of other types of diabetes or glucose intolerance	N in B	13741	7895	2904	n.e.	n.e.	5782	n.e.	78
	% of B in A	1.1%	1.8%	1.5%	-	-	0.3%	-	1.7%
	PRR if B added	+4.3%	+8.3%	+3.5%	-	-	+6.8%	-	+1.5%
DIAG_T2DM_PC OR **DIAG_T2DM_INP** OR **DIAG_T2DM_SC** OR **DIAG_DMUNSPEC** OR **DIAG_DMUNSPEC_OTH**	N in B	191999	43622	13098	253197	62191	79035	779	2520
	% of B in A	73.2%	56.2%	15.1%	100.0%	92.7%	94.3%	100.0%	100.0%
	PRR if B added	+2.4%	+0.0%	+7.6%	+0.0%	+0.0%	+2.5%	+0.0%	+2.2%
DRUG_INSULIN ≥2 prescriptions/ dispensings of insulin in one calendar year	N in B	45522	22074	21192	41019	15020	11607	n.e.	n.e.
	% of B in A	17.9%	25.4%	25.8%	16.1%	19.0%	12.3%	-	-
	PRR if B added	+0.0%	+3.0%	+10.9%	+0.1%	+3.4%	+2.0%	-	-
DRUG_INSULIN_ONE ≥1 prescriptions/ dispensings of insulin	N in B	62341	23319	0	0	17719	0	18	0
	% of B in A	21.2%	26.5%	-	-	22.0%	-	1.5%	-
	PRR if B added	+3.4%	+3.6%	-	-	+4.4%	-	+0.8%	-
DRUG_ORAL ≥2 prescriptions/ dispensings of NIAD in one calendar year	N in B	216338	57153	57712	136370	51589	45624	-	0
	% of B in A	85.2%	71.0%	100.0%	51.7%	76.9%	53.0%	-	-
	PRR if B added	+0.0%	+2.7%	+0.0%	+2.1%	+0.0%	+2.9%	-	-
DRUG_ORAL_ONE ≥1 prescriptions/ dispensings of NIAD	N in B	273952	61604	0	0	54181	62110	45	0
	% of B in A	87.5%	72.7%	-	-	80.8%	70.6%	5.8%	-
	PRR if B added	+20.3%	+6.7%	-	-	+0.0%	+5.4%	+0.0%	-
DRUG_INSULIN OR **DRUG_INSULIN_ONE** OR **DRUG_ORAL** OR **DRUG_ORAL_ONE**	N in B	295676	70405	64016	151576	58355	65076	40	0
	% of B in A	93.0%	81.1%	100.0%	57.7%	82.6%	73.1%	50.0%	-
	PRR if B added	+23.4%	+9.6%	+10.9%	+2.2%	+4.4%	+6.6%	+4.1%	-

Table 3. Continued

TEST_TEST_GLUCO5_1YR ≥5 glycated hemoglobin tests in 1 year	N in B	266940	16999	0	0	0	0	0	0
	% of B in A	45.8%	21.6%	-	-	-	-	-	-
	PRR if B added	+59.3%	+0.3%	-	-	-	-	-	-
TEST_GLUCO2_PYEAR_5YRS ≥2 glycated hemoglobin tests per year during 5 consecutive years	N in B	172784	28583	0	0	0	0	0	0
	% of B in A	32.6%	36.1%	-	-	-	-	-	-
	PRR if B added	+35.4%	+0.7%	-	-	-	-	-	-
TEST_GLUCO5_1YR OR **TEST_GLUCO2_PYEAR_5YRS**	N in B	335466	34801	0	0	0	0	0	0
	% of B in A	52.8%	44.1%	-	-	-	-	-	-
	PRR if B added	+79.2%	+0.8%	-	-	-	-	-	-
LABVAL_FAST_GLUC ≥2 fasting glucose values >126mg/dl	N in B	0	0	0	0	0	32153	0	0
	% of B in A	-	-	-	-	-	38.6%	-	-
	PR if B added	-	-	-	-	-	+0.8%	-	-
LABVAL_HbA1c ≥2 glycated hemoglobin value >6.5%	N in B	0	0	62400	0	44271	20196	0	0
	% of B in A	-	-	65.1%	-	63.6%	24.1%	-	-
	PRR if B added	-	-	+43.0%	-	+2.4%	+0.7%	-	-
LABVAL_LCURVE_GLUC ≥2 glucose tolerance test values >200mg/dl	N in B	0	0	0	0	0	32	0	0
	% of B in A	-	-	-	-	-	0.0%	-	-
	PRR if B added	-	-	-	-	-	+0.0%	-	-
LABAL_FAST_GLUC OR **LABAL_HbA1c** OR **LABAL_LCURVE_GLUC**	N in B	0	0	62400	0	44271	38764	0	0
	% of B in A	-	-	65.1%	-	63.6%	46.5	-	-
	PRR if B added	-	-	+43.0%	-	+2.4%	+1.0%	-	-

Since patients can be identified by more than one component algorithms, percentages may overlap.
Grey cells correspond to component algorithms that were included in the relevant recommended composite algorithm.
NIAD: Non-Insulin Antidiabetic Drugs.
A= recommended composite algorithm.
B= tested component algorithm(s).
N= Study population.
PRR= prevalence rate ratio of "A or B" in N with respect to the percentage of A in N.

In RLDs, DRUG-based components identified from 81% to 100% of the respective total case population, while from 58% to 83% in PCDs. TEST-based components were included in the recommended composite algorithm of the Danish RLD only in which these algorithms identified 44% of the total case population. Although TEST-based components were also extracted from the Italian RLD, they were not included in the recommended composite algorithm since they would have almost doubled the total case population (PRR=+79.2%), thus suggesting a too low specificity. LABVAL-based algorithms were included in the recommended composite algorithm of the Italian PCD only: overall, the three components from this data domain identified 46% of the total case population. Notably, subjects from the same data source could be identified by one or more component thus the percentages reported above may overlap.

DISCUSSION

Through the application of the standard data derivation procedure tested in this study, cases of T2DM were identified in eight distinct sources of health data with heterogeneous characteristics. Logical combinations of standardized component algorithms, each based on a single data domain, were used to build data source-tailored case-finding algorithms. This "component algorithm strategy" facilitated both benchmarking and interpretation of results across data sources. It also allowed the assessment of the impact of individual standardized component algorithms on the total population of cases retrieved in each participating data source that ultimately provided insight into the strengths and limitations of each data source with respect to the identification of T2DM cases.

Compared to previous projects that aimed to combine different European sources of EHR for research purposes,[2, 3] the main innovation of the standard procedure tested in this study was the use of component algorithms as building blocks that could be combined to create more complex case-finding algorithms. As demonstrated by the results presented here, in the context of a multi-national, multi-data source study, the "component algorithm strategy" represents an extremely flexible approach for generating EHR-driven[6] case-finding algorithms in a standardized fashion: on the one hand, it allows the local experts' knowledge of the EHR "natural system"[8] to be fully leveraged, avoiding

loss of information and assuring the correctness of the derived information, while, on the other hand, it facilitates the interpretation and benchmarking of results obtained even across data sources with very different characteristics. Notably, the data derivation procedure tested in this study requires that all component algorithms locally available for the identification of the condition of interest should be extracted, tested and stored regardless whether they will be subsequently included in the final recommended composite algorithm. This also gives to investigators and local experts the chance to tweak the preferred identification algorithm at the study design stage, according to the study questions.

Gaining insight into cases identified by data source-tailored case-finding algorithms
In this study, the composite algorithms recommended by local experts for the identification of T2DM were extremely variable, resulting, however, in a selection of cases that are likely to represent the best possible local approximation of the true case identification. Indeed, since the age-specific prevalence of diabetes is expected to be fairly homogeneous across the geographic areas we are considering,[32] the observed differences in terms of percentage of the corresponding study populations can be interpreted in light of both the specific components adopted and of relevant data sources' characteristics. Among all data sources, the highest percentage of cases was identified in HD because this data source only captures subjects who visit the hospital, who, by definition, will have a higher burden of disease with respect to the general population. On the other extreme, the BD showed the lowest percentage, possibly because people volunteering to participate in this data source are slightly healthier than the general population. Both HD and BD identify patients with T2DM using DIAG-based component only. However, while in HD cases were identified among inpatients only who are expected to be at a more advanced stage of the disease and more likely to have comorbidities,[5] in BD characteristics of cases were probably more representative of patients with T2DM in the corresponding source population, because diagnoses are recorded in a primary care setting. As for the three primary care data sources, the Italian PCD adopted a case finding strategy based on data from DIAG and LABVAL. This strategy was expected to be very sensitive. Moreover, in a previous validation study, it was also proven to have the highest possible PPV. [30] Therefore, its recommended algorithm can be considered an excellent approximation of a true case identification and

the observed percentage of cases can be assumed to be a valid estimate of the prevalence of T2DM in the correspondent source population. In the PCD from UK a lower percentage of cases was identified compared to the Italian PCD. This result could be due to a slight underreporting of diagnoses in the data source. As for the Dutch PCD, the age-specific percentage of detected cases was almost identical to that observed in the PCD from UK. However, in the Dutch PCD a DRUG-based algorithm was adopted as additional inclusion criterion to the DIAG-based component DIAG_T2DM_PC, since the latter was not sensitive enough when used alone. In fact, general practitioners participating to the Dutch PCD often record diagnoses using free text description which may sometimes remain elusive to the keywords-based retrieval process. Among RLDs, the percentage of the population identified in the Dutch RLD was slightly lower than that observed in the other two RLDs from Italy and Denmark respectively. Indeed, local experts of the Dutch RLD recommended the use of one single DRUG-based component (DRUG_ORAL) as the preferred case-finding algorithm, while the other two data sources, on the grounds of previous validation studies,[27, 31] adopted more complex composite algorithms that allowed to increase sensitivity by including also components based on DIAG and/or TEST. In particular, the Danish RLD was the only data source collecting diagnoses from secondary care. Notably, TEST-based components, which identify patients through specific patterns of utilization of glycated haemoglobin tests, were not included in the Italian RLD since they resulted to be far more unspecific than in the Danish RLD. This was clearly showed when the impact of TEST-based components on the total population of cases identified in the two data sources was observed. Such a difference was probably due to local healthcare system organization and guidelines with respect to diagnosis and follow-up of diabetic patients.

Understanding quality of a local case-finding algorithm

In studies utilizing routinely collected health data, understanding the quality of local case-finding algorithms is paramount for the interpretation of study findings[11, 33] and *a fortiori* in multi-data source studies. The component algorithm strategy proposed in this study can indirectly provide approximation of algorithm validity indexes, even when no formal validation studies are available for one or more of the participating data sources. This is attained through the benchmarking of components and composite algorithms across

data sources with similar characteristics but collecting data from different geographic areas or *vice versa*.

Indeed, in this study, cases in PCDs were basically identified through primary care diagnoses and are thus expected to be fairly representative of the T2DM patients in the corresponding source populations. In RLDs, instead, most of cases were captured through non-insulin antidiabetic drugs utilization which cannot identify those patients at a earlier stage of the disease who are not on drug treatment (do diet only) and may also misclassify T2DM with other diseases for which the same drugs can sometimes be used (e.g. polycystic ovary syndrome).[4] Supposing that the validity of the latter case-finding algorithm was completely unknown, data reported in Table 3 can be used to obtain an approximation of its expected sensitivity and PPV. As an example, the Dutch RLD, which used a case-finding algorithm based on the utilization of non-insulin antidiabetics only (i.e. DRUG_ORAL) can be considered. Since sensitivity corresponds to the percentage of subjects with a true diagnosis of T2DM who also have the DRUG_ORAL pattern of non-insulin antidiabetic drugs utilization, this percentage can be estimated from the Dutch PCD to be around 77%, or slightly lower if we accept that sensitivity in the Dutch PCD is not 100% (the corresponding percentage in the other two PCD data sources is lower than 55%). PPV, instead, is the percentage of subjects utilizing oral antidiabetics who really have type 2 diabetes. In this case, value higher than 90% is expected since other indications for such drugs have a very low prevalence. [4] In fact, this is also confirmed in both PCDs from Italy and UK where the component DRUG_ORAL added less than 3% of cases when used as additional inclusion criteria.

Tailoring selection of components to a research question

Investigators and local experts could consider changing their preferred identification algorithm according to the type of study question or sensitivity analysis: in the case of a study involving T2DM, for instance, if specificity is important, they may switch to DIAG-based algorithms at the expenses of sensitivity; if sensitivity is important, they may add other inclusion criteria, like TEST-based components; if homogeneity across different data sources is important, they may agree to adopt a DRUG-based strategy.

Limitations

Although this was a proof-of-concept study in which results obtained were not intended as estimates of disease frequencies, limitations that might have biased the results and comparisons discussed in this manuscript must be acknowledged. In particular, validation of the retrieved cases was not performed as well as important variables other than age were not considered for stratification of results.

CONCLUSIONS

Through the identification of T2DM cases, this study demonstrates that the standard procedure for data derivation developed within the EMIF project represent a methodological advancement for the execution of multi-national, multi-data source studies. In fact, on the basis of a shared definition of any event of interest, the procedure assures interoperability of heterogeneous EHR systems and allows establishing data-source tailored case-identification algorithm in a standardized fashion, providing sufficient information for contextualization and correct interpretation of study results and generating transparent and reusable documentation on the entire data derivation process.

REFERENCES

1 Richesson RL, Horvath MM, Rusincovitch SA. Clinical research informatics and electronic health record data. Yearb Med Inform 2014;9(1):215-23.

2 Coloma PM, Schuemie MJ, Trifiro G, Gini R, Herings R, Hippisley-Cox J, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol Drug Saf 2011 Jan;20(1):1-11.

3 Avillach P, Coloma PM, Gini R, Schuemie M, Mougin F, Dufour JC, et al. Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. J Am Med Inform Assoc 2013 Jan 1;20(1):184-92.

4 Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN, Miranda ML, et al. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc 2013 Dec;20(e2):e319-e326.

5 Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. PLoS One 2014;9(11):e110900.

6 Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc 2013 Dec;20(e2):e206-e211.

7 Overby CL, Pathak J, Gottesman O, Haerian K, Perotte A, Murphy S, et al. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. J Am Med Inform Assoc 2013 Dec;20(e2):e243-e252.

8 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013 Jan 1;20(1):117-21.

9 Valkhoff VE, Coloma PM, Masclee GM, Gini R, Innocenti F, Lapi F, et al. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. J Clin Epidemiol 2014 Aug;67(8):921-31.

10 Gini R, Schuemie M, Brown J, Ryan P. Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research: A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies. eGEMs 2016;4(1, Article 2).

11 Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med 2015 Oct;12(10):e1001885.

12 Conway M, Berg RL, Carrell D, Denny JC, Kho AN, Kullo IJ, et al. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. AMIA Annu Symp Proc 2011;2011:274-83.

13 Trifiro G, Coloma PM, Rijnbeek PR, Romio S, Mosseveld B, Weibel D, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? J Intern Med 2014 Jun;275(6):551-61.

14 Ryden L, Grant PJ, Anker SD, Berne C, Cosentino F, Danchin N, et al. ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD: the Task Force on diabetes, pre-diabetes, and cardiovascular diseases of the European Society of Cardiology (ESC) and developed in collaboration with the European Association for the Study of Diabetes (EASD). Eur Heart J 2013 Oct;34(39):3035-87.

15 Gini R, Francesconi P, Mazzaglia G, Cricelli I, Pasqua A, Gallina P, et al. Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey. BMC Public Health 2013;13:15.

16 Vlug AE, van der LJ, Mosseveld BM, van Wijk MA, van der Linden PD, Sturkenboom MC, et al. Postmarketing surveillance based on electronic patient records: the IPCI project. Methods Inf Med 1999 Dec;38(4-5):339-44.

17 Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Inform Prim Care 2011;19(4):251-5.

18 Nexo BA, Pedersen L, Sorensen HT, Koch-Henriksen N. Treatment of HIV and risk of multiple sclerosis. Epidemiology 2013 Mar;24(2):331-2.

19 Johannesdottir SA, Horvath-Puho E, Ehrenstein V, Schmidt M, Pedersen L, Sorensen HT. Existing data sources for clinical epidemiology: The Danish National Database of Reimbursed Prescriptions. Clin Epidemiol 2012;4:303-13.

20 Herk-Sukel MP, Lemmens VE, Poll-Franse LV, Herings RM, Coebergh JW. Record linkage for pharmacoepidemiological studies in cancer patients. Pharmacoepidemiol Drug Saf 2012 Jan;21(1):94-103.

21 Mayer MA, Furlong LI, Torre P, Planas I, Cots F, Izquierdo E, et al. Reuse of EHRs to Support Clinical Research in a Hospital of Reference. Stud Health Technol Inform 2015;210:224-6.

22 Sancho JJ, Planas I, Domenech D, Martin-Baranera M, Palau J, Sanz F. IMASIS. A multicenter hospital information system--experience in Barcelona. Stud Health Technol Inform 1998;56:35-42.

23 Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol 2015 Aug;44(4):1137-47.

24 Ryden L, Grant PJ, Anker SD, Berne C, Cosentino F, Danchin N, et al. ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD: the Task Force on diabetes, pre-diabetes, and cardiovascular diseases of the European Society of Cardiology (ESC) and developed in collaboration with the European Association for the Study of Diabetes (EASD). Eur Heart J 2013 Oct;34(39):3035-87.

25 World Health Organization. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus. Geneva; 1999. Report No.: WHO/NCD/NCS/99.2.

26 Ryden L, Grant PJ, Anker SD, Berne C, Cosentino F, Danchin N, et al. ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD: the Task Force on diabetes, pre-diabetes, and cardiovascular diseases of the European Society of Cardiology (ESC) and developed in collaboration with the European Association for the Study of Diabetes (EASD). Eur Heart J 2013 Oct;34(39):3035-87.

27 Carstensen B, Kristensen JK, Ottosen P, Borch-Johnsen K. The Danish National Diabetes Register: trends in incidence, prevalence and mortality. Diabetologia 2008 Dec;51(12):2187-96.

28 Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc 2012 Mar;19(2):212-8.

29 Ryden L, Grant PJ, Anker SD, Berne C, Cosentino F, Danchin N, et al. ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD: the Task Force on diabetes, pre-diabetes, and cardiovascular diseases of the European Society of Cardiology (ESC) and developed in collaboration with the European Association for the Study of Diabetes (EASD). Eur Heart J 2013 Oct;34(39):3035-87.

30 Gini R, Schuemie MJ, Mazzaglia G, Lapi F, Pasqua A, Dazzi P, et al. Automatic identification of stages of type 2 diabetes, hypertension, ischaemic heart disease and heart failure from Italian General Practioners' electronic medical records: a validation study [Abstract]. Pharmacoepidemiol Drug Saf Sep 2015;24:1-587 2015.

31 Gini R, Schuemie MJ, Mazzaglia G, Lapi F, Francesconi P, Pasqua A, et al. Identifying chronic conditions from data sources with incomplete diagnostic information: the case of Italian administrative databases [Abstract]. Pharmacoepidemiol Drug Saf Sep 2015;24:1-587 2015.

32 Prevalence estimates of diabetes, adults aged 20-79 years, 2011. OECD-iLibrary 2011Available from: URL: http://www.oecd-ilibrary.org/sites/9789264183896-en/01/14/g1-14-01.html?itemId=/content/chapter/9789264183896-17-en&_csp_=bd0a209a2ba2bc37c7e2885 1b7808bdc

33 Hernan MA. With great data comes great responsibility: publishing comparative effectiveness research in epidemiology. Epidemiology 2011 May;22(3):290-1.

GENERAL DISCUSSION

MAIN FINDINGS

The main objective of this thesis was advancing the methodology of validation studies of case-finding algorithms that exploit diversity across available data, rather than collecting new data. The main case study was whether and how Italian administrative databases can be used to identify cases with chronic diseases and to monitor standards of care, using medical records as a reference data source. In the second part of this thesis we described how this validation approach compares with other initiatives in the world, and how it applies to European networks of diverse data sources.

Part I: chronic diseases and compliance with standards of care in Italian Administrative Databases

Part I was devoted to address the following question: what are the optimal algorithms to detect chronic diseases in Italian Administrative Databases (IAD), and what is the validity of estimates of compliance with standards of care? The work is described in five chapters with the following specific questions

1. How do the prevalence estimates derived from finding cases of chronic diseases in IAD compare with estimates derived from other data sources?
2. What is the validity of algorithms detecting chronic diseases and their level of severity from medical records of General Practitioners?
3. What are the optimal case-finding algorithms in IAD to find cases of chronic diseases?
4. How do estimates of compliance with standards of care derived from IAD compare with estimates derived from the Health Search (HSD), a database of medical records of the Italian College of General Practitioners?
5. How do measures of compliance with standards of care derived from IAD compare with measures derived from the medical records of the General Practitioners?

At first in chapter 1 we demonstrated that IAD data could be used to identify four pre-specified chronic diseases by using codes and utilization of care patterns. The prevalence estimates were consistently lower than population estimates

from HSD and surveys, for diabetes, ischaemic heart disease and heart failure, but the geographic pattern was the same in all data sources. We suspected that IAD data may not have perfect sensitivity and specificity.

To verify the assumption that medical records collected by GPs in HSD correctly identify chronic diseases, we assessed the positive predictive value (PPV) of case-finding algorithms in primary care medical records in chapter 2. The answer was reassuring for type 2 diabetes, hypertension and ischaemic heart disease, but not for heart failure: the PPV was only 55%. Prevalence estimates from HSD were higher than prevalence estimated from the yearly national survey, as we had noticed in chapter 1, which suggests that the false negative rate was low.

Based on these results we assumed that primary care medical records were providing a gold standard for type 2 diabetes, hypertension and ischaemic heart disease. This was the key assumption we needed to address question 3, which dealt with finding the optimal queries for these chronic diseases in IAD. In chapter 3 we tested dozens of case-finding algorithms for type 2 diabetes, hypertension and ischaemic heart disease in IAD, using linked primary care medical records as a gold standard. In order to structure this process we first listed a set of simple *component algorithms*: discharge diagnosis from hospitalizations, diagnosis from exemptions (an Italian registry of disease-specific exemptions from copayment of health care), drug utilization, utilization of diagnostic services. We tested each of them separately, then combined them in more complex strategies: components were used as inclusion, exclusion or refinement criteria. For each disease we identified the composed algorithm with the best balance between sensitivity and positive predictive value. The resulting optimal algorithms all had high positive predictive value, but low sensitivity. In particular for ischaemic heart disease sensitivity was lower than 50%. Moreover, we were able to identify several strategies that may be chosen for studies where sensitivity is more important than specificity: drug utilization or utilization of diagnostic services is an important component of such algorithms. Vice versa, if high positive predictive value is paramount for a study, algorithms based on diagnostic components (inpatient and exemptions) could be considered, at least for sensitivity analysis.

Validity of compliance with standards of care for chronic diseases in Italian Administrative Databases (questions 4 and 5)

Based on the case finding algorithms from chapter 1, we compared estimates of compliance with standards of care (diagnostic tests and recommended drug

treatment) between IAD and HSD at the same geographical level (chapter 4), for ischaemic heart disease estimates, diabetes and heart failure. The comparison was reasonable. This was especially true for compliance with recommended therapies, whilst diagnostic tests were underreported in IAD. However, since we could not link the IAD and HSD data at the individual level, we doubted whether the observed similarities in chapter 4 were true, or a coincidence of a mixture of confounding factors. Linkage of the data at the individual level between IAD and medical records in HSD showed that the latter was true (chapter 5). We saw that the numerators of therapy indicators were consistently measured across data sources, while the numerators of diagnostic indicators were not. Moreover, persons in the IAD denominator had lower compliance with respect to the false negatives (persons with the disease but who are not included in the IAD denominator). These two opposite effects resulted in a substantial similarity between estimates from IAD and medical records, except for some of the diagnostic indicators. So compliance with standards measured with IAD for the diseases of interest compared well with measures from medical records, but comparability seemed to be the result of coincidence rather than real similarity.

Part II: deriving study variables for multi-database studies

The main research question of Part II was: how do networks of databases generate study variables in the different sites, and how do they assess variable validity?

We split the question in two parts.

6. How is the derivation of study variables organized in existing data networks in Europe and in the United States?
7. How can a network of diverse data sources in Europe streamline the process of data derivation?

Data derivation in existing networks

In chapter 6 we described derivation of study variables in 4 different networks: in Italy (MATRICE), Europe (EU-ADR) and in the United States (Mini-Sentinel and OMOP).

In Mini-Sentinel the focus was on data from different sources in the United States, which have a mild internal diversity: the coding system was the same (ICD9CM), and most of the participant nodes were providing administrative data. The derivation of study variables was conducted with the simplest algorithm: detecting the occurrence of a code in a set of ICD9CM diagnostic codes. Sensitivity of this approach was discussed, and alternatives were described, by systematically reviewing available literature. When literature review was considered insufficient, the positive predictive value of diagnostic codes was measured using chart review as a gold standard. Validity indices were not corrected for in the statistical analysis.

In the OMOP network several algorithms based on systematic literature reviews were used to identify the same study variable. Coding systems for diagnosis, drugs and diagnostic tests were not equal across data partners, but were mapped to a same set of vocabularies altogether, independently on the study variable of interest. Instead of validating case-finding algorithms, an indirect form of evaluation was conducted: a set of *known* associations between the study variable and exposure to various drugs was collected from literature and regulatory sources. The quality of the algorithms for a study variable was identified with the ability to tell true from false associations.

The MATRICE network used the automatic queries on medical records validated in chapter 2 to validate its study variables, and had the possibility of testing dozens of combinations of component algorithms and explicitly estimate validity indices.

In the European EU-ADR network the heterogeneity of original data sources was considerably higher with respect to the country-based networks, as it covered different countries health care structures and data base structures. The overall process to derive study variables was handled as an iterative process, where local data experts were providing their feedback to central proposals, and was protocol-based. Terminology mapping from different coding systems was addressed per single study variable. As for the provenance (primary, secondary, inpatient care or death) of the records selected in each local data source, and additional inclusion, exclusion and refinement criteria, they were discussed before decisions were made, but the process was never embedded in accessible documents. Validation was performed for two diseases.

A data derivation workflow for heterogeneous data sources: the component algorithm strategy

Based on the experiences in MATRICE and the EU-ADR project we proposed a data derivation workflow for the EMIF Project, a European network of diverse data sources (chapter 7). The main step forward was the explicit use of the component algorithm strategy, which originated from the validation study in Italy (chapter 3).

We formalized the notion of *component algorithm*: in this wider context this is an algorithm which selects data from a single data domain (diagnosis, drugs, diagnostic tests, results from diagnostic tests), and specifies the provenance of the records (primary care, secondary care, inpatient care, death, other) as well as the pattern used to identify the case (at least one record, at least two records in a specified time frame, …) and the choice of the case date (date of the first record, date of the second record…). Each person is positive to one, more than one, or none of a list of components.

As a case study, we aimed to identify the best site-tailored case-finding algorithm for type 2 diabetes mellitus in 8 heterogeneous European data sources. We created a list of 17 components. In each data source, all the persons were labelled as positive or negative per component. Population frequency of each combination of components was counted and the resulting aggregated data could be shared. Each data source chose a tailored combination of components as their preferred case-finding algorithm: this approach combined standardization and flexibility, and fully documented the local choices. We could establish sensitivity and PPV of the component considering non-insulin antidiabetic drug utilization *within HSD*. Sensitivity resulted to be 55%, similar to the 59% sensitivity *in IAD* that we had estimated in the individual-level study. In the EMIF network we had a similar pair of databases from the Netherlands: a primary care database, and an administrative database. By analogy, we assumed that the sensitivity of the same drug component from the Dutch administrative database was similar to 77%, which was the sensitivity of the same component in the Dutch primary care database. So we could provide an estimate of sensitivity, using only aggregated data.

This way we could *exploit* diversity across sites, as a mean to infer validity. Rather than considering it a weakness, we demonstrated that it can indeed be considered a strength.

GENERALIZABILITY OF FINDINGS

Monitoring quality of healthcare for chronic diseases in Italy

Our validation studies covered a limited number of Italian regions and were conducted in a specific point in time, for a few diseases. We discuss here whether the results from the validation studies can be generalized.

Generalization of validity of case-finding algorithms to other chronic diseases, over time and for other Italian regions.
In IAD we saw some general patterns in sensitivity and positive predictive value of component algorithms, in different diseases and over regions. Such general patterns may support hypothesis generation with respect to the validity for other diseases, or of the same diseases over time. In the following we discuss this, component per component.
Validity indices of components based on inpatient and exemption diagnosis were similar across the three diseases: they all had high positive predictive value and low sensitivity, although the specific measure varied both by disease and by region. The three conditions had in common a relatively straightforward clinical definition. If we considered other chronic diseases whose diagnostic algorithm is more complex, as is the case for chronic obstructive pulmonary disease [Romanelli2016], positive predictive value of inpatient exemption diagnoses may be lower. We saw in chapter 2 that ambiguity in the clinical definition of heart failure compromised significantly the positive predictive value of a case-finding algorithm from primary care medical records: the same is expected to happen for components based on inpatient and exemption diagnosis. It is not expected that positive predictive value changes significantly over time, unless new diagnostic criteria or tests become available. On the contrary, sensitivity may be affected by reduced use of inpatient care, but the effect could only be seen in the long run, if new cases are less often hospitalized.
As for drug utilization components, the positive predictive value is associated with prevalence of other conditions which are indications for the same drug. Non-insulin antidiabetic drugs have type 2 diabetes as the largely most common indication, therefore positive predictive value was very high. It is expected to be similarly high, for instance, for memantine or cholinesterase inhibitors as components of a case-finding algorithm for dementia, because these drugs have

basically no other indications. On the contrary, it is likely to be low in the case of patterns of utilization of respiratory drugs as with component algorithms for chronic obstructive pulmonary disease, as we saw in chapter 1, because asthma is a highly prevalent concurrent indication [Gini2015, Romanelli2016]. Sensitivity of a drug utilization component is associated with the strength and penetration of the recommendation to treat the disease, as well as to the accessibility of the drug: this was very high for antithrombotic therapy in ischaemic heart disease, but would be lower for dementia [Francesconi2007]. Validity of drug-based components change over time, due to the entrance of new drugs in the market, new indications (possibly off-label) for old drugs, new clinical guidelines and regulatory interventions.

Components based on utilization of diagnostic tests had promising results in terms of sensitivity. This family of components is commonly used in the Danish administrative databases, which have a similar structure in this respect to IAD [Carstensen2008]. This type of components is expected to provide a relevant contribution in the near future, when the data become available on the whole Italian population. However, access to this type of healthcare services in the healthcare system may change over time, according to new rules in copayment and reduced accessibility.

Our estimates relied on a sample of local health units from 5 of the 21 Italian regions. We adopted a convenience sampling schema, and regions in the South were under represented. The determinants we identified for geographical variation of validity indices (for instance, local off-label use of drugs) should be discussed with local experts when generalizing our estimates.

In summary, some patterns in the positive predictive value and sensitivity that we observed may be generalized over time, over regions, and potentially provide guidance for validity of case-finding algorithms of other diseases. However, we don't recommend to generalize the estimates that we obtained for validity indices to other diseases. Hypotheses on validity should be generated and tested against external data sources, using the methodology that we will explain in the subsections below.

Generalization of the validity of measures of compliance with standards of care for chronic diseases in Italy with Italian Administrative Databases

We demonstrated that we should be cautious when measuring compliance with standards of care for type 2 diabetes, hypertension and ischaemic heart disease from IAD. The balance between the number of false negatives and their reduced compliance, as well as access to diagnostic services contracted by the healthcare services, may change over time and between regions, and hamper the accuracy of estimates.

At the population level, drug utilization is consistently recorded by IAD and medical records: this is likely to be the case for other drugs which are reimbursed by the healthcare system. In the case of compliance with recommended therapies for a chronic disease, in general, the obstacle to validity of IAD measures amounts essentially to the quality of the denominator (see previous subsection), and to the different level of compliance in false negatives. The profile of false negatives is specific per disease: diagnosis recorded in inpatient care detect more severe patients, who are likely to be more adherent to medications; drug utilization components detect a population which is more likely to be adherent to any drug. Generalizability of the results of chapter 5 to this respect is limited: to understand the expected validity of the measure of compliance with a new standard, all those aspects should be evaluated by a panel of specialist and primary care physicians, hypothesis should be generated, and a new round of validation should be performed.

In the case of compliance with diagnostic recommendations, an additional level of complexity is due to uneven accessibility to diagnostic services. Interpretation of results and generalizability to other diseases, should be carefully discussed with local managers of the healthcare system.

In the next subsection a generalization of the methodology we developed in this thesis for validation studies on existing databases is discussed. This approach could be applied routinely to support interpretation of measures of compliance to standards of care for chronic diseases in IAD.

GENERALIZABILITY OF METHODOLOGY

A general approach for individual-level validation studies on existing databases
In the first part of this thesis we used one data source to systematically validate another, at the individual patient level. We now draw the main lessons from this specific experience, to propose a general methodological approach.

In its essence, the starting point to apply our approach is the following

- two data sources are available on the same population, but record linkage at the individual level between them is not possible on a regular and continuous basis
- in one of the data sources ("worse" data source) the optimal case-finding algorithm for a study variable is unknown, while in the other ("better" data source) a good or excellent case-finding algorithm is available
- permission for record linkage on a sample of persons can be obtained for purposes of validation

The approach for an individual-level validation of the variable of interest comprises three steps

Ecological-level comparison: compare at the ecological level the population prevalence (or incidence) of the variable, obtained separately from the two data sources; compare the resulting estimates with each other and with estimates available from the literature or from other sources (survey data, disease registries): the rationale for this step is obtaining evidence that the quality of the two data sources is as expected, before engaging in more complex studies

Gold standard validation: validate the candidate case-finding algorithm in the "better" data source with manual chart review, on a small sample, to confirm its validity: the rationale of this step is confirming the quality of the "better" data source, using a traditional methodology

Individual-level validation: obtain permission to perform the record-linkage between the two data sources, and use the validated case-finding algorithm on the "better" data source as a gold standard to find the optimal case-finding algorithm in the "worse" data source. If the case-finding algorithm on the "better" data source is proven by the "Gold standard validation" step to have imperfect validity, it can still be used to provide information on the other data source, but statistical adjustment needs to be applied [Gart1966] (see section "Implications for research" below).

As we saw in the previous subsections, the results of our validation studies cannot be generalized to other diseases, and may cease to be reliable over time. What is generalizable is the approach that we just described. This has the advantage of all the big data studies: fast, repeatable, reproducible on vast populations.

Generalization to multi-database studies: component analysis

We consider now the situation when a study variable must be defined in multiple data sources, that are considered part of the same study. The steps described in the previous subsection can be used systematically, although with some modifications. Indeed, in a multi-database setting there are some limitations:

- data sources may not be available on *the same* population, but rather on *similar* populations;
- even if two data sources are available on the same population, individual-level record-linkage is unlikely to be allowed, due to time or legal constraints

On the positive side, in multi-database studies more than two data sources are normally available.

To fully exploit the potential of a multi-database setting, it is proposed to generalize the component algorithm strategy. A set of component algorithms must be created, each specified with data domain, provenance of data, pattern of records, date, and length of look-back. Sets of aggregated data like the one represented in Table 1 can be shared among partners. Multiple comparisons across data sources can be performed, as described in the remaining subsections. This process is referred to as *component analysis*. At the end of the analysis, a data source-tailored combination of components can be chosen as the preferred case-finding algorithm, depending on the characteristics of the data source, on the results from the component analysis, and on the requirements of the study protocol.

Analysis of components in a same data source

If a component is a gold standard (on the basis of a local validation study), then validity of the other components in the same data source can be obtained from direct calculation (generalization of the case of the drug utilization component in chapter 7)

If a component has known validity (on the basis of a local validation study), then validity of the other components in the same data source can be obtained from statistical estimation, assuming conditional independence of the two components [Gart1966]

Inference of validity of a component from one data source to another
Validity of drug utilization components can sometimes be generalized from one source to another. We have seen that utilization of some drugs in Italy was very consistently measured by primary care medical records and by administrative data. This is likely to be the case for all drugs which are reimbursed by the national health care system in Italy, but do not need a specialist prescription. This happens not only in Italy, but also in other countries where General Practitioners have a "gatekeeper" role and take-over specialist prescriptions. The observation that medical records and administrative data, measure drug utilization in a consistent manner was made recently in a very general setting [Hripcsak2016]. Under the assumption that data utilization is measured consistently by different data sources in the same population, validity of a drug utilization component in that population is independent on the data source where it is measured. If in a data source a "gold standard" algorithm is available, then validity of drug utilization can be estimated. Based on the previous remark, validity of a drug utilization component is independent on the data source where it is measured, so we can make the assumption that the validity of the drug utilization component is similar in other data sources in the same population.
If validity can be estimated this way in several regions, and it results in similar estimates, it can be argued that validity is similar even in other regions. Indeed, ultimately, sensitivity of drug utilization is the percentage of patients with a disease who are treated, and positive predictive value is the percentage of treated patients who are treated for that indication. Both those percentages are influenced by national and international clinical guidelines and regulatory interventions - even though local circumstances may have a relevant impact (as we have seen in chapter 4).

Analysis of prevalence or incidence of the same component across different data sources

Prevalence or incidence of the same component can be compared across different data sources: for instance, rates of hospitalized acute events, or prevalence of chronic diseases diagnoses in primary care, or drug utilization prevalence. Observed differences can be discussed. They may be attributed to population differences, or to differences in the underlying health care system, or in granularity of the coding system. If no such difference is plausible, we can infer that the explanation lies in difference in validity of the variable.

Estimates of positive predictive value of a component and of prevalence of the disease provide an estimate of sensitivity

If positive predictive value of a component is estimated (for instance, from a validation study, or based on assumptions), then the expected number of false positives can be computed from the number of persons identified by the component. The number of true positives can be obtained by difference. If population prevalence of the disease can be obtained from an external data source, then the number of persons with the disease in the data source can be estimated. An estimate of the sensitivity is the ratio of true positives to the estimated number of persons with the disease.

Table 1. Template of a dataset of aggregated data for component analysis, in the case of 3 components and 2 data sources.

DATA SOURCE	COMPONENT 1	COMPONENT 2	COMPONENT 3	Number of subjects
A	No	No	No	
A	Yes	No	No	
A	No	Yes	No	
A	No	No	Yes	
A	Yes	Yes	No	
A	Yes	No	Yes	
A	Yes	Yes	No	
A	Yes	Yes	Yes	
B	No	No	No	
B	Yes	No	No	
B	No	Yes	No	
B	No	No	Yes	
B	Yes	Yes	No	
B	Yes	No	Yes	
B	Yes	Yes	No	
B	Yes	Yes	Yes	

IMPLICATIONS FOR RESEARCH

Research needed to improve estimates of compliance with standards of care for chronic diseases based on Italian Administrative Databases

We recommend to be cautious when measuring compliance with standards of care for chronic diseases from IAD. A key development would be assessing whether a calibration of the estimates from IAD improves the estimates of compliance with standards of care. To do so, a statistical model should be built that incorporates validity indices of the denominators and concordance between estimates of the numerators. The result should be compared with the "best estimate".

Research needed to improve validation studies on existing databases

In the absence of a true gold standard, a reference standard can still be used to validate another algorithm, but statistical adjustment needs to be applied. This is a classical approach, and relies on the assumption of conditional independence between the reference standard and the algorithm to be validated [Gart1966]. This assumption is unlikely to be met in our setting, and research is needed to explore how it can be mitigated.

As for multi-database studies, the component analysis we proposed in the "Generalizability of findings" section would benefit from further elaboration on several theoretical issues. First, robustness of validity estimates to violations of the assumptions needs to be assessed. Second, effort should be made to adapt the classical Hui-Walter paradigm, which was developed to validate diagnostic tests without a gold standard, to the case-finding algorithm situation. In this approach, the true disease status is modelled as a latent class, and information derived from applying the same test to populations with different prevalence is exploited to estimate the parameters of the model [Hui1980, Joseph1995]. In an evolution of this approach, a similar model can be estimated when several tests are applied to the same population [Walter1988, Johnson2001]. The Hui-Walter paradigm relies on assumptions that are unlikely to be met when case-finding algorithms are involved instead of diagnostic tests, so research is needed to understand how they can be mitigated. Third, in order to implement components across data sources which use different coding systems, mapping between different terminologies must be handled. Mapping must take place

within multiple data domains: diagnosis, drugs, diagnostic tests, interventional procedures. We saw that two different strategies were adopted in different networks: either terminology mapping was done once and for all, or it was tuned on a specific study variable. The first strategy requires a bigger effort of initial mapping, but later allows for homogeneous treatment of mapped data across data sources. The second is conceivably producing study variables with higher validity indices. Understanding determinants of validity loss would be a relevant achievement.

RECOMMENDATIONS

Monitoring quality of healthcare for chronic diseases in Italy

Measures of compliance with standards of care for chronic diseases obtained from IAD should be used with caution. For the measures studied in this thesis the following recommendations on their use can be made:

- pay-for-performance schemes for general practitioners should at present not be based on such measures. The validity issues are still too large and using them at this stage for reimbursement purposes would most likely introduce perverse incentives.
- At regional level, IAD measures can be used alongside measures from primary care medical records for quality audits. The measures are informative when taking the local context into account.
- At national level, when presenting comparisons between regions an open dialogue is essential and the following should be considered:
    - interpretation should be facilitated by presenting estimates of percentages of false negatives and estimates of regional use of private services, using as a reference
        - validation studies from this thesis,
        - analysis of HSD
        - analysis of national surveys
    - regional managers should be encouraged to comment on the data, using insights derived from local audits as a basis
- At national level, if compliance with standards of care is used to *evaluate* regions, calibrated estimates of true compliance should be built
- When using compliance with standards of care as an outcome in policy evaluation studies based on IAD, the impact of misclassification should be assessed or discussed.

As an additional recommendation, validation studies should be routinely performed on clusters of patients. This recommendation was recently endorsed in a report of the Italian National Institute of Health [ISS2014]. The results from such studies could be used to update validity parameters, using the methods developed in this thesis.

Overall the compliance measures should be "handled with care", but when applied with knowledge of the context and limitations of the underlying data they can provide meaningfull information to assess and improve the quality of care.

Deriving variables in multi-database studies

When designing an observational study in a multi-database network, we recommend to adopt a component algorithm strategy, as described in the section "Generalizability of findings" above, to define the key study variables. In a feasibility phase, we encourage data sources to share aggregated data to allow estimates of validity indices. The combination of components meant to define a study variable should be tailored to each data source, in order to maximize validity. At the design stage, sensitivity analysis using different components to define variables should be designed, to assess robustness of study results with respect to heterogeneity in validity across sites. Including validity indices at the analytical step should be considered as well.

FROM BIG DATA TO LOCAL INTELLIGENCE

We saw in this thesis how big data can be used to validate big data, and we recommend to exploit systematically this opportunity to enhance local intelligence. This is paramount to meet the expectations from policymakers, clinicians, patients and citizens, that big data will provide more detailed and reliable guidance to the choices they need to make, in order to improve health and health care.

REFERENCES

[Carstensen2008] Carstensen B, Kristensen JK, Ottosen P, Borch-Johnsen K. The Danish National Diabetes Register: trends in incidence, prevalence and mortality. Diabetologia 2008 Dec;51(12):2187-96.

[Gart1966] Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. Am J Epidemiol. 1966;83(3):593–602.

[Francesconi2007] Francesconi P, Gini R, Roti L, Bartolacci S, Corsi A, Buiatti E. The Tuscany experimental registry for Alzheimer's disease and other dementias: how many demented people does it capture? Aging Clin Exp Res. 2007;19(5):390–3.

[Gini2015] Gini R, Di Domenicantonio R, Ferroni E, Pasqua A, Roberto G, Martigli Jiang M, et al. [Identify patients with chronic obstructive pulmonary disease from Italian Administrative Databases: the MATRICE-BPCO study of Agenas in four Italian regions.] In Italian. XXXIX conference of the Italian Epidemiology Association. Poster available at https://www.ars.toscana.it/files/progetti/malattie_croniche/MATRICE-BPCO/poster_gini_AIE_MATRICE-BPCO.pdf

[Hripcsak2016] Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proceedings of the National Academy of Sciences. 2016;201510502.

[Hui1980] Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics. 1980;36(1):167–71.

[ISS2014] Istituto Superiore di Sanità. Rapporto Registri e Sorveglianze. Pubblicazione dell'Istituto Superiore di Sanità. 14/23 Pt 1. 2014. Available at http://www.iss.it/binary/publ/cont/14_23_pt_1_web.pdf

[Johnson2001] Johnson WO. Screening without a «Gold Standard»: The Hui-Walter Paradigm Revisited. American Journal of Epidemiology. 2001;153(9):921–4.

[Joseph1995] Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol. 1995;141(3):263–72.

[Romanelli2016] Romanelli AM, Raciti M, Protti MA, Prediletto R, Fornai E, Faustini A. How Reliable Are Current Data for Assessing the Actual Prevalence of Chronic Obstructive Pulmonary Disease? PLoS One. 2016;11(2).

[Walter1988] Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J Clin Epidemiol. 1988;41(9):923–37.

SUMMARY/SAMENVATTING

BACKGROUND AND RESEARCH QUESTIONS

The era of big data opens new means to improve health and health care. Observational evidence from large populations can provide guidance to the choices of multiple stakeholders. Policymakers can understand how health care systems can be better organized, clinicians can explore more in detail all the treatment options, patients can put peculiarities of their own diseases at the centre of the clinical decisions, and citizens can obtain evidence to inform their political options. New methodologies need to be developed, and traditional tools and ways of thinking need to be renovated to adapt to the new perspective, to avoid drawing incorrect conclusions from studies based on this resource [Ray2011, Mooney 2015].

The methodological challenge relies not only in the observational nature of the studies that can be conducted on big data, but also in the heterogeneous characteristics of the data that is now available. We are interested in the latter.

The objective of a validation study of a case-finding algorithm is estimating the *validity indices* of the algorithm, that quantify to which extent the study variable corresponds to the true variable.

The methodology for validation studies has been developed in the context of validation of *diagnostic tests*, that is, procedures collecting clinical parameters from patients [Whiting2003]. The parameters of sensitivity and specificity of a diagnostic test are rooted in human biology and essentially depend on the characteristics of the population where they are estimated [Greenland1996]. On the contrary, validity indices of a case-finding algorithm in a database depend on multiple characteristics of the system, for instance completeness of data collection, accuracy in coding habits, granularity of the coding system, organization of the health care system in the geographic area where the data is collected – besides characteristics of the population whose data is collected. All those factors are subject to change from one database to another, and over time [Quan2009, Reich2012, Herret2013, Morley2014, Rahimi2014, Lanes2015]. Generalizing estimates of validity parameters of a case-finding algorithm outside of the environment where a validation study was executed is questionable.

In order to support effectively the interpretation of the results of a study, validation of study variables should be as close as possible to the actual dataset that is used for the statistical analysis. Traditional validation studies, which imply manual assessment of samples of records, are time-consuming and expensive [Hernan2011]. An alternative solution is *exploiting big data to validate big data*: this methodological advancement is emerging in the recent literature.

The main objective of this thesis is advancing the methodology of validation studies of case-finding algorithms that exploit diversity across available data, rather than collecting new data.

The case study that led us to this advancement was the assessment of the capacity of the Italian administrative database to capture cases of chronic disease to get estimates for the compliance with standards of care. Primary care medical records were the main comparative source. Part I of this thesis is focussed on this topic.
In Part II we exploited the results and extended the methodology of Part I to the context of multi-database, multi-national studies.

Validation of variables defining chronic disease and compliance with standards of care in Italian Administrative Databases
Italy has a universal, single-payer healthcare system. Chronic diseases impose an increasing burden on the Italian aging population, and are a major threat to sustainability of the healthcare system [OECD2015].
Administrative data are collected on a large set of services provided to the population, and are available for secondary analysis to health policy makers. Secondary use of Italian Administrative Databases (IAD) to detect patients with chronic conditions would allow surveillance, planning, monitoring of quality of healthcare, as well as assessment of impact of new organizational models on relevant health and quality outcomes.

The MATRICE Project, funded by the Italian Ministry of Health, was launched in 2011 by the Italian National Agency for Regional Healthcare Services (AGENAS), with the aim of defining methodologies and tools to best exploit administrative data for the purposes of monitoring quality of healthcare for patients with chronic diseases.

The main research question of Part I of the thesis was: what are the optimal algorithms to detect chronic diseases in the IAD, and what is the validity of estimates of compliance with standards of care?

We split this main question in 5 specific questions.

1. How do the prevalence estimates derived from finding cases of chronic diseases in IAD compare with estimates derived from other data sources?
2. What is the validity of algorithms detecting chronic diseases and their level of severity from medical records of the General Practitioners?
3. What are the optimal case-finding algorithms in IAD to find cases of chronic diseases?
4. How do estimates of compliance with standards of care derived from IAD compare with estimates derived from the Health Search (HSD), a database of medical records of the Italian College of General Practitioners?
5. How do measures of compliance with standards of care derived from IAD compare with measures derived from the medical records of the General Practitioners?

Validity of variables in multi-database studies

In 2004, after five year of widespread use, the anti-inflammatory drug rofecoxib was withdrawn from the market due to severe safety concerns. It was estimated that if a monitoring system had been in place querying the medical records of 100 million patients, the adverse cardiovascular effect would have been discovered in just few months. After this episode, networks of researchers with regular access to observational healthcare data sources have been created. Methods and procedures have been generated to execute studies in a distributed fashion, to take advantage both from size and from diversity of the populations they could merge [Trifiro2014]. Those advantages come at the price that the level of complexity in deriving study variables scales up, because all the characteristics that have an impact on validity may be different across sites.

Table 1. Summary of the studies in Part I of this thesis. IAD: Italian Administrative Databases.

Variable to be validated	Study design	Data sources	Diseases	Setting	Year	Chapter
Case-finding algorithm	Ecological comparison between existing data	IAD, primary care medical records, national survey	Ischaemic heart disease, diabetes mellitus, heart failure, chronic obstructive pulmonary disease	Population samples from 5 Italian regions: Veneto, Emilia Romagna, Tuscany, Marche and Sicily	2009	1
Case-finding algorithm	Individual-level manual validation	Primary care medical records, manual assessment	Type 2 diabetes, hypertension, ischaemic heart disease and heart failure, with levels of severity	300 cases per disease, from 12 GPs across Italy	2014	3
Case-finding algorithm	Individual-level record-linkage between existing data	IAD, primary care medical records	Type 2 diabetes, hypertension, ischaemic heart disease and heart failure	25 clusters of subjects, 5 per each of the following regions: Lombardy, Veneto, Emilia Romagna, Tuscany, Puglia	2012	4
Compliance with standards of care	Ecological comparison between existing data	IAD, primary care medical records	Ischaemic heart disease, diabetes mellitus, heart failure	Population samples from 5 Italian regions: Veneto, Emilia Romagna, Tuscany, Marche and Sicily	2009	2
Compliance with standards of care	Individual-level record-linkage between existing data	IAD, primary care medical records	Type 2 diabetes, hypertension and ischaemic heart disease	25 clusters of subjects, 5 per each of the following regions: Lombardy, Veneto, Emilia Romagna, Tuscany, Puglia	2012	5

This is especially true in Europe: diversity in local mechanisms of data collection is a consequence of the diversity among European countries in language, culture, political and health care organization. Notwithstanding, cross-border evidence is necessary to address common questions such as efficacy and safety of medical products and vaccines, or comparison of quality of health care.

Part II of this thesis was devoted to investigate the process adopted by some existing networks to derive the study variables in each site and address their validity, and to propose and test a novel methodology to streamline this process.

Research questions

The main research question of Part II was: how do networks of databases handle the process of generating study variables in the different sites, and how do they assess their validity?

We split the question in two parts.

6. How is the derivation of study variables organized in existing data networks in Europe and in the United States?
7. How can a network of diverse data sources in Europe streamline the process of data derivation?

MAIN FINDINGS

Chronic diseases and compliance with standards of care in Italian Administrative Databases

At first in chapter 1 we demonstrated that IAD data could be used to identify four pre-specified chronic diseases by using codes and utilization of care patterns. The prevalence estimates were consistently lower than population estimates from HSD and surveys, for diabetes, ischaemic heart disease and heart failure, but the geographic pattern was the same in all data sources. We suspected that IAD data may not have perfect sensitivity and specificity.

To verify the assumption that medical records collected by GPs in HSD correctly identify chronic diseases, we assessed the positive predictive value (PPV) of case-finding algorithms in primary care medical records in chapter 2. The answer was reassuring for type 2 diabetes, hypertension and ischaemic heart disease. Prevalence estimates from HSD were higher than prevalence estimated from the yearly national survey, as we had noticed in chapter 1, which suggests that the false negative rate was low.

Based on these results we assumed that primary care medical records were providing a gold standard. This was the key assumption we needed to address question 3. In chapter 3 we tested dozens of case-finding algorithms for type 2 diabetes, hypertension and ischaemic heart disease in IAD, using linked primary care medical records as a gold standard. In order to structure this process we first listed a set of simple *component algorithms*: discharge diagnosis from hospitalizations, diagnosis from exemptions, drug utilization, utilization of

diagnostic services. We tested each of them separately, then combined them in more complex strategies. The resulting optimal algorithms all had high positive predictive value, but low sensitivity.

Based on the case finding algorithms from chapter 1, we compared estimates of compliance with standards of care (diagnostic tests and recommended drug treatment) between IAD and HSD at the same geographical level (chapter 4), for ischaemic heart disease, diabetes and heart failure. The comparison was reasonable. This was especially true for compliance with recommended therapies, whilst diagnostic tests were underreported in IAD. However since we could not link the IAD and HSD data at the individual level, we doubted whether the observed similarities in chapter 4 were true, or a coincidence of a mixture of confounding factors. Linkage of the data at the individual level between IAD and medical records in HSD showed that the latter was true (chapter 5). Compliance with standards measured with IAD for the diseases of interest compared well with measures from medical records, but comparability seemed to be the result of coincidence rather than real similarity.

Deriving study variables for multi-database studies

In chapter 6 we described derivation of study variables in 4 different networks: in Italy (MATRICE), Europe (EU-ADR) and in the United States (Mini-Sentinel and OMOP).

In Mini-Sentinel the focus was on data from different sources in the United States, which have a mild internal diversity. The derivation of study variables was conducted with the simplest algorithm: detecting the occurrence of a code in a set of ICD9CM diagnostic codes. When literature review was considered insufficient, the positive predictive value of diagnostic codes was measured using chart review as a gold standard. Validity indices were not corrected for in the statistical analysis.

In the OMOP network several algorithms based on systematic literature reviews were used to identify the same study variable. Coding systems for diagnosis, drugs and diagnostic tests were not equal across data partners, but were mapped to a same set of vocabularies altogether, independently on the study variable of interest. The quality of the algorithms for a study variable was identified with the ability to tell true from false associations in a predefined set of drug-outcome pairs.

The MATRICE network used the automatic queries on medical records validated in chapter 2 to validate its study variables, and had the possibility of testing dozens of combinations of component algorithms and explicitly estimate validity indices.
In the European EU-ADR network the heterogeneity of original data sources was considerably higher with respect to the country-based networks, as it covered different countries, health care structures and data base structures. The overall process to derive study variables was protocol-based. Terminology mapping from different coding systems was addressed per single study variable. As for the provenance (primary, secondary, inpatient care or death) of the records selected in each local data, the process was never embedded in accessible documents. Validation was performed for two diseases.

A data derivation workflow for heterogeneous data sources: the component algorithm strategy
Based on the experiences in MATRICE and the EU-ADR project we proposed a data derivation workflow for the EMIF Project, a European network of diverse data sources (chapter 7). The main step forward was the explicit use of the component algorithm strategy, which originated from the validation study in Italy (chapter 3).
We formalized the notion of *component algorithm*: in this wider context this is an algorithm which selects data from a single data domain (diagnosis, drugs, diagnostic tests, results from diagnostic tests), and specifies the provenance of the records (primary care, secondary care, inpatient care, death, other) as well as the pattern used to identify the case (at least one record, at least two records in a specified time frame, …) and the choice of the case date (date of the first record, date of the second record…). Each person is positive to one, more than one, or none of a list of components.
As a case study, we aimed to identify the best site-tailored case-finding algorithm for type 2 diabetes mellitus in 8 heterogeneous European data sources. We created a list of 17 components. In each data source, all the persons were labelled as positive or negative per component. Population frequency of each combination of components was counted and the resulting aggregated data could be shared. Each data source chose a tailored combination of components as their preferred case-finding algorithm: this approach combined standardization and flexibility, and fully documented

the local choices. Moreover, we could provide an estimate of sensitivity of a component, using only aggregated data.

This way we could *exploit* diversity across sites, as a mean to infer validity. Rather than considering it a weakness, we demonstrated that it can indeed be considered a strength.

Generalizability of findings

In IAD we saw some general patterns in sensitivity and positive predictive value of component algorithms, in different diseases and over regions, and could potentially provide guidance for validity of case-finding algorithms of other diseases. However we don't recommend to generalize the estimates that we obtained for validity indices to other diseases. Hypotheses on validity should be generated and tested against external data sources, using the methodology that we will explain below.

We demonstrated that we should be cautious when measuring compliance with standards of care for type 2 diabetes, hypertension and ischaemic heart disease from IAD. The balance between the number of false negatives and their reduced compliance, as well as access to diagnostic services contracted by the healthcare services, may change over time and between regions, and hamper the accuracy of estimates. Interpretation of results and generalizability to other diseases, should be carefully discussed with local managers of the healthcare system.

What is generalizable is the approach that we followed. This approach could be applied routinely to support interpretation of measures of compliance to standards of care for chronic diseases in IAD.

In its essence, the starting point to apply our approach is the following

- two data sources are available on the same population, but record linkage at the individual level between them is not possible on a regular and continuous basis

- in one of the data sources ("worse" data source) the optimal case-finding algorithm for a study variable is unknown, while in the other ("better" data source) a good or excellent case-finding algorithm is available
- permission for record linkage on a sample of persons can be obtained for purposes of validation

The approach for an individual-level validation of the variable of interest comprises three steps

Ecological-level comparison: compare at the ecological level the population prevalence of the variable, obtained separately from the two data sources; compare the resulting estimates with each other and with estimates available from the literature or from other sources, to obtain evidence that the quality of the two data sources is as expected

Gold standard validation: validate the candidate case-finding algorithm in the "better" data source with manual chart review

Individual-level validation: obtain permission to perform the record-linkage between the two data sources, and use the validated case-finding algorithm on the "better" data source as a gold standard to find the optimal case-finding algorithm in the "worse" data source.

We consider now the case when a study variable must be defined in several data sources, who want to collaborate in a same study. The approach described above can be used systematically, although with some modifications. Indeed, in a multi-database setting we have some limitations:

- data sources may not be available on *the same* population, but rather on *similar* populations;
- even if two data sources are available on the same population, individual-level record-linkage is unlikely to be allowed, due to time or legal constraints

On the positive side, in multi-database studies more than two data sources are normally available.

We proposed to generalize the *component algorithm strategy*. A set of component algorithms must be created, each specified with data domain, provenance of data, pattern of records, date, and length of look-back. Sets of aggregated data can be shared among partners. Multiple comparisons across data sources can be performed.

- Analysis of components in a same data source

- Inference of validity of a component from one data source to another
- Analysis of prevalence or incidence of the same component across different data sources
- Estimates of positive predictive value of a component and of prevalence of the disease provide an estimate of sensitivity

We refer to this process as *component analysis*. At the end of the analysis, a data source-tailored combination of components can be chosen as the preferred case-finding algorithm, depending on the characteristics of the data source, on the results from the component analysis, and on the requirements of the study protocol.

Implications for research

We recommend to be cautious when measuring compliance with standards of care for chronic diseases from IAD. A key development would be assessing whether a calibration of the estimates from IAD improves the estimates of compliance with standards of care. To do so, a statistical model should be built that incorporates validity indices of the denominators and concordance between estimates of the numerators. The result should be compared with the "best estimate".

In the absence of a true gold standard, a reference standard can still be used to validate another algorithm, but statistical adjustment needs to be applied. This is a classical approach, and relies on the assumption of conditional independence between the reference standard and the algorithm to be validated [Gart1966]. This assumption is unlikely to be met in our setting, and research is needed to explore how it can be weakened.

As for multi-database studies, the component analysis would benefit from further elaboration on several theoretical issues. First, robustness of validity estimates to violations of the assumptions needs to be assessed. Second, effort should be made to adapt the classical Hui-Walter paradigm, which was developed to validate diagnostic tests without a gold standard, to the case-finding algorithm situation [Hui1980, Walter1988, Joseph1995, Johnson2001]. Third, in order to implement components across data sources which use different coding systems,

mapping between different terminologies must be handled. Mapping must take place within multiple data domains: diagnosis, drugs, diagnostic tests, interventional procedures. We saw that two different strategies were adopted in different networks: either terminology mapping was done once and for all, or it was tuned on a specific study variable. The first strategy requires a bigger effort of initial mapping, but later allows for homogeneous treatment of mapped data across data sources. The second is conceivably producing study variables with higher validity indices. Understanding determinants of validity loss would be a relevant achievement.

Recommendations

Monitoring quality of healthcare for chronic diseases in Italy

Measures of compliance with standards of care for chronic diseases obtained from IAD should be used with caution. For the measures studied in this thesis the following recommendations on their use can be made:

- pay-for-performance schemes for general practitioners should at present not be based on such measures. The validity issues are still too large and using them at this stage for reimbursement purposes would most likely introduce perverse incentives.
- At regional level, IAD measures can be used alongside measures from primary care medical records for quality audits. The measures are informative when taking the local context into account.
- At national level, when presenting comparisons between regions an open dialogue is essential and the following should be considered:
    - interpretation should be facilitated by presenting estimates of percentages of false negatives and estimates of regional use of private services, using as a reference
        - validation studies from this thesis,
        - analysis of HSD
        - analysis of national surveys
    - regional managers should be encouraged to comment on the data, using insights derived from local audits as a basis
- At national level, if compliance with standards of care is used to *evaluate* regions, calibrated estimates of true compliance should be built

- When using compliance with standards of care as an outcome in policy evaluation studies based on IAD, the impact of misclassification should be assessed or discussed.

As an additional recommendation, validation studies should be routinely performed on clusters of patients. This recommendation was recently endorsed in a report of the Italian National Institute of Health [ISS2014]. The results from such studies could be used to update validity parameters, using the methods developed in this thesis.

Overall the compliance measures should be "handled with care", but when applied with knowledge of the context and limitations of the underlying data they can provide meaningfull information to assess and improve the quality of care.

Deriving variables in multi-database studies

When designing an observational study in a multi-database network, we recommend to adopt a component algorithm strategy to define the key study variables. In a feasibility phase, we encourage data sources to share aggregated data to allow estimates of validity indices. The combination of components meant to define a study variable should be tailored to each data source, in order to maximize validity. At the design stage, sensitivity analysis using different components to define variables should be designed, to assess robustness of study results with respect to heterogeneity in validity across sites. Including validity indices at the analytical step should be considered as well.

From big data to local intelligence

We saw in this thesis how big data can be used to validate big data, and we recommend to exploit systematically this opportunity to enhance local intelligence. This is paramount to meet the expectations from policymakers, clinicians, patients and citizens, that big data will provide more detailed and reliable guidance to the choices they need to make, in order to improve health and health care.

REFERENCES

[Altman1994] Altman DG, Bland JM. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. BMJ. 1994;308(6943):1552.

[Altman1994b] Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994;309(6947):102.

[AMD-SID2014] Italian Diabetologist Association, Italian Society for Diabetology. [Italian standards for treatment of diabetes mellitus] [In Italian]. http://www.standarditaliani.it/skin/www.standarditaliani.it/pdf/STANDARD_2014_May28.pdf

[Amed2011] Amed S, Vanderloo SE, Metzger D, Collet J -P, Reimer K, McCrea P, et al. Validation of diabetes case definitions using administrative claims data. Diabetic Medicine. 2011;28(4):424–7.

[Benchimol2011] Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. Journal of Clinical Epidemiology. 2011;64(8):821–9.

[Benchimol2015] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLOS Medicine. 2015;12(10):e1001885.

[Bowker2015] Bowker SL, Savu A, Lam NK, Johnson JA, Kaul P. Validation of administrative data case definitions for gestational diabetes mellitus. Diabet Med. 2015

[Carnahan2012] Carnahan RM. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data: summary of findings and suggestions for future research. Pharmacoepidemiology and Drug Safety. 2012;21:90–9.

[Carstensen2008] Carstensen B, Kristensen JK, Ottosen P, Borch-Johnsen K. The Danish National Diabetes Register: trends in incidence, prevalence and mortality. Diabetologia 2008 Dec;51(12):2187-96.

[Cutrona2013] Cutrona SL, Toh S, Iyer A, Foy S, Daniel GW, Nair VP, et al. Validation of acute myocardial infarction in the Food and Drug Administration's Mini-Sentinel program. Pharmacoepidemiol Drug Saf. 2013;22(1):40–54.

[Ferretti2009] Ferretti S, Guzzinati S, Zambon P, Manneschi G, Crocetti E, Falcini F, et al. [Cancer incidence estimation by hospital discharge flow as compared with cancer registries data] [In Italian]. Epidemiol Prev.2009;33(4–5):147–53.

[Francesconi2007] Francesconi P, Gini R, Roti L, Bartolacci S, Corsi A, Buiatti E. The Tuscany experimental registry for Alzheimer's disease and other dementias: how many demented people does it capture? Aging Clin Exp Res. 2007;19(5):390–3.

[GarciaRodriguez2010] García Rodríguez LA, Ruigómez A. Case validation in research using large databases. Br J Gen Pract. 1 2010;60(572):160–1.

[Gart1966] Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. Am J Epidemiol. 1966;83(3):593–602.

[Gini2015] Gini R, Di Domenicantonio R, Ferroni E, Pasqua A, Roberto G, Martigli Jiang M, et al. [Identify patients with chronic obstructive pulmonary disease from Italian Administrative Databases: the MATRICE-BPCO study of Agenas in four Italian regions.] In Italian. XXXIX conference of the Italian Epidemiology Association. Poster available at https://www.ars.toscana.it/files/progetti/malattie_croniche/MATRICE-BPCO/poster_gini_AIE_MATRICE-BPCO.pdf

[Gorina2011] Gorina Y, Kramarow EA. Identifying Chronic Conditions in Medicare Claims Data: Evaluating the Chronic Condition Data Warehouse Algorithm. Health Services Research. 2011;46(5):1610–27.

[Greenland1996] Greenland S. Basic methods for sensitivity analysis of biases. Int J Epidemiol. 1996;25(6):1107–16.

[Hernan2011] Hernan MA. With great data comes great responsibility: publishing comparative effectiveness research in EPIDEMIOLOGY. Epidemiology. 2011;22(3):290–1.

[Herret2013] Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, Staa T van, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. BMJ. 2013;346:f2350.

[Hripcsak2016] Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proceedings of the National Academy of Sciences. 2016;201510502.

[Hui1980] Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics. 1980;36(1):167–71.

[ISS2014] Istituto Superiore di Sanità. Rapporto Registri e Sorveglianze. Pubblicazione dell'Istituto Superiore di Sanità. 14/23 Pt 1. 2014. Available at http://www.iss.it/binary/publ/cont/14_23_pt_1_web.pdf

[John2016] John A, McGregor J, Fone D, Dunstan F, Cornish R, Lyons RA, et al. Case-finding for common mental disorders of anxiety and depression in primary care: an external validation of routinely collected data. BMC Med Inform Decis Mak. 2016

[Johnson2001] Johnson WO. Screening without a «Gold Standard»: The Hui-Walter Paradigm Revisited. American Journal of Epidemiology. 2001;153(9):921–4.

[Joseph1995] Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol. 1995;141(3):263–72.

[Lanes2015] Lanes S, Brown JS, Haynes K, Pollack MF, Walker AM. Identifying health outcomes in healthcare databases. Pharmacoepidemiol Drug Saf. 2015;24(10):1009–16.

[Lanes2015b] Lanes S, Esposito D. Case ascertainment in electronic data bases: Where's Waldo? Community Meeting at Innovation in Medical Evidence Development and Surveillance (IMEDS). July 2015. http://imeds.reaganudall.org/sites/default/files/IMEDS_Outline_Dv8.pdf

[Kahn2010] Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. Br J Gen Pract. 2010;60(572):e128–36.

[Lix2008] Lix LM, Yogendran MS, Shaw SY, Burchill C, Metge C, Bond R. Population-based data sources for chronic disease surveillance. Chronic Dis Can. 2008;29(1):31–8.

[Lix2008b] Lix LM, Yogendran MS, Leslie WD, Shaw SY, Baumgartner R, Bowman C, et al. Using multiple data features improved the validity of osteoporosis case ascertainment from administrative databases. Journal of Clinical Epidemiology. 2008;61(12):1250–60.

[Mooney2015] Mooney SJ, Westreich DJ, El-Sayed AM. Commentary: epidemiology in the era of big data. Epidemiology. 2015;26(3):390–4.

[Morley2014] Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation. PLoS ONE. 2014;9(11):e110900.

[Nosyk2013] Nosyk B, Colley G, Yip B, Chan K, Heath K, Lima VD, et al. Application and Validation of Case-Finding Algorithms for Identifying Individuals with Human Immunodeficiency Virus from Administrative Data in British Columbia, Canada. Medeiros R, curatore. PLoS ONE. 2013;8(1):e54416.

[OECD2015] OECD. OECD Reviews of Health Care Quality. Italy 2014: Raising Standards. OECD Publishing; 2015.

[Quan2009] Quan H, Khan N, Hemmelgarn BR, Tu K, Chen G, Campbell N, et al. Validation of a Case Definition to Define Hypertension Using Administrative Data. Hypertension. 2009;54(6):1423–8.

[Quantin2013] Quantin C, Benzenine E, Velten M, Huet F, Farrington CP, Tubert-Bitter P. Self-controlled case series and misclassification bias induced by case selection from administrative hospital databases: application to febrile convulsions in pediatric vaccine pharmacoepidemiology. Am J Epidemiol. 2013;178(12):1731–9.

[Rahimi2014] Rahimi A, Liaw S-T, Taggart J, Ray P, Yu H. Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in Electronic Health Records. International Journal of Medical Informatics. 2014;83(10):768–78.

[Ray2011] Ray WA. Improving Automated Database Studies: Epidemiology. 2011;22(3):302–4.

[Reich2012] Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inform. 2012;45(4):689–96.

[Romanelli2016] Romanelli AM, Raciti M, Protti MA, Prediletto R, Fornai E, Faustini A. How Reliable Are Current Data for Assessing the Actual Prevalence of Chronic Obstructive Pulmonary Disease? PLoS One. 2016;11(2).

[Trifiro2014] Trifirò G, Coloma PM, Rijnbeek PR, Romio S, Mosseveld B, Weibel D, et al. Combining multiple healthcare databases for post-marketing drug and vaccine safety surveillance: why and how? J Intern Med. 2014

[Valkhoff2014] Valkhoff VE, Coloma PM, Masclee GMC, Gini R, Innocenti F, Lapi F, et al. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. Journal of Clinical Epidemiology. 2014;67(8):921–31.

[Walter1988] Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J Clin Epidemiol. 1988;41(9):923–37.

[Whiting2003] Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Medical Research Methodology. 2003;3:25.

SAMENVATTING

Achtergrond en onderzoeksvragen

Het tijdperk van 'big data' opent tal van nieuwe mogelijkheden om de gezondheidszorg te laten leren en verbeteren door hergebruik van de data die juist in de zorg worden gegenereerd. Observationele data van grote populaties kunnen de keuzes van de verschillende belanghebbenden, richting geven. Beleidsmakers kunnen uit deze data bijvoorbeeld opmaken hoe de gezondheidszorg beter georganiseerd kan worden, clinici kunnen behandelingsopties beter verkennen en patiënten kunnen zichzelf vergelijken met andere patiënten met dezelfde ziekte. Nieuwe methodologieën moeten worden ontwikkeld, en de traditionele gereedschappen en denkwijzen moeten worden aangepast naar dit nieuwe perspectief. Belangrijk is dat er geen verkeerde conclusies worden getrokken op basis van het gebruik van deze bronnen [Ray2011, Mooney 2015]. De methodologische uitdagingengaan niet alleen over de methoden van onderzoek maar vooral ook de heterogeniteit van de data die nu beschikbaar zijn en komen. In dit proefschrift waren we vooral geïnteresseerd in het laatste.

De belangrijkste doelstelling van dit proefschrift is het bevorderen van validatie studies voor het identificeren van 'ziekten' (case finding algoritmes) op een wijze dat combinatie van data in verschillende systemen gebruikt wordt en geen nieuwe data verzameld hoeven te worden.

In dit proefschrift kijken we vooral naar de mogelijkheden van de Italiaanse administratieve databases (IAD) om gevallen van chronische ziektente identificeren, wat nodig is om de kwaliteit van zorg beter te kunnen monitoren. Hierbij werden de data van de IAD vergeleken met elektronische medische dossiers van een huisartsen database in Italië. Deel I van dit proefschrift is gericht op dit onderwerp. In deel II wordt de methodologie die in deel I is beschreven verder uitgebreid naar de context van multinationale database studies.

Validatie van variabelen definiëren chronische ziekte en naleving van de normen van de zorg in de Italiaanse administratieve databanken

Het Matrice Project, gefinancierd door het Italiaanse ministerie van Volksgezondheid, werd in 2011 gelanceerd door het Italiaanse Nationaal Agentschap voor regionale Healthcare Services (AGENAS), met als doel

methoden en instrumenten te ontwikkelen die gebruikt zouden kunnen worden ten behoeve van het monitoren van de kwaliteit van de gezondheidszorg voor patiënten met chronische aandoeningen.
De centrale onderzoeksvraag in het eerste deel van het proefschrift is: wat zijn de beste algoritmen om chronische ziekten op te sporen in IAD, en wat is de validiteit van de resultaten voor monitoring van de kwaliteit van zorg?
In dit proefschrift hebben we bewezen dat de berekeningen over het voorkomen van enkele chronische ziekten in IAD goed overeenkomen met de aanwezigheid van ziekte op populatie en individueel niveau in huisartsen gegevens in Italië. Een lijst van eenvoudige component algoritmen (diagnose van hospitalisaties, de diagnose van vrijstellingen, geneesmiddel gebruik, het gebruik van diagnostische testen) moet altijd worden getest om te kijken wat de impact is op de berekeningen. De optimale algoritmes hadden hogevoorspellende waarde, maar een lage gevoeligheid.
Wij adviseren om voorzichtig te zijn bij het meten van de kwaliteit van de zorg van chronische ziekten indien gebruik gemaakt wordt van IAD. Ze moeten worden gebruikt met inachtneming van de context en de beperkingen van de onderliggende gegevens waarna ze kunnen bijdragen aan het monitoren van de kwaliteit van zorg.

Geldigheid van variabelen in Multi-database-studies
In 2004, werd het anti-inflammatoire geneesmiddelen rofecoxib uit de markt genomen vanwege een verhoogd risico op myocard infarct, dit was vijf jaar na introductie van het middel en na massaal gebruik in de Verenigde Staten. Deze affaire leidde tot hervorming van het vigilantie systeem in de USA en door Richard Plattwerd geponeerd dat als er een systeem zou zijn metde anonieme medische dossiersvan 100 miljoen patiëntende nadelige cardiovasculaire effecten in enkele maanden ontdekt zouden zijn. Deze discussies waren het begin van de FDA Amendment Act (FDA-AA). Deze verplicht de FDA om gedistribueerde data van tenminste 100 miljoen mensen te gebruiken voor het monitoren van de veiligheid van geneesmiddelen. Dit is nu gerealiseerd in het Sentinel systeem. Parallel hieraan werden in Europa en Canada methoden en procedures ontwikkeld om studies uit te voeren op een gedistribueerde manier. Hierdoor kan nu worden geprofiteerd van zowel de grootte als van de diversiteit van verschillende bevolkingen [Trifiro2014]. Het combineren van data is vooral in Europa niet makkelijk omdat elk land zijn eigen gezondheidszorg systeem,

coderingen en databases heeft. Er is echter een duidelijke noodzaak om algemene vragen, zoals de werkzaamheid en veiligheid van geneesmiddelen en vaccins, of de vergelijking van de kwaliteit van de gezondheidszorg samen aan te pakken. In deel II van dit proefschrift beschrijven we het data transformatie proces in 4 verschillende netwerken: in Italië (Matrice), Europa (EU-ADR) en in de Verenigde Staten (Mini-Sentinel en OMOP). In Mini-Sentinellag de nadruk op de data van verschillende bronnen in de Verenigde Staten, die slechts matig verschillen. In het OMOP netwerk werden vooral methoden ontwikkeld om snel verschillende designs te implementeren voor signaal detectie. In het Europese Unie ADR-netwerk is de heterogeniteit van de oorspronkelijke gegevensbronnen aanzienlijk hoger dan in de nationale netwerken. In dit netwerk wordt niet de hele database omgezet in een geharmoniseerd model maar wordt protocol gericht gewerkt.

In het EMIF Project hebben we het concept en de methode voor algoritme ontwikkeling op basis van componenten die werd gestart in Matrice, toegepast voor type 2 diabetes mellitus in 8 Europese heterogene gegevensbronnen. We bevelen aan dit als algemene strategie toe te passen in gedistribueerde netwerken. De component analyse zou gebaat zijn bij verdere uitwerking van verschillende theoretische vraagstukken. We zagen in dit proefschrift hoe gegevens in de ene databron kunnen worden gebruikt om grote hoeveelheden data in andere bronnen te valideren. Door de hoeveelheden data die beschikbaar komen in de big data era' zou de methodologie en de mogelijkheden hiervan beter ontwikkeld moeten worden. Dit is van cruciaal belang om aan de verwachtingen van beleidsmakers, artsen, patiënten en burgers te voldoen en uiteindelijk de gezondheid en de gezondheidszorg te verbeteren.

REFERENCES

[Mooney2015] Mooney SJ, Westreich DJ, El-Sayed AM. Commentary: epidemiology in the era of big data. Epidemiology. 2015;26(3):390–4.

[Ray2011] Ray WA. Improving Automated Database Studies: Epidemiology. 2011;22(3):302–4.

[Trifiro2014] Trifirò G, Coloma PM, Rijnbeek PR, Romio S, Mosseveld B, Weibel D, et al. Combining multiple healthcare databases for post-marketing drug and vaccine safety surveillance: why and how? J Intern Med. 2014

ABOUT THE AUTHOR

Rosa Gini was born in Italy in 1970. In 1995 she graduated with a degree in Mathematics in Milan, and started a PhD in Pisa. In 1997 and 1999 she gave birth to Eloisa and Sara, with her colleague and partner, Maurizio Parton. In 1999 she dreamt the proof of her first theorem and in 2001 she defended her PhD thesis in Differential Topology. Over the next two years, she enjoyed doing some more mathematics in the field of Differential Geometry, while lifting from Maurizio some of his parental duties in order for him to pursue his academic career, in the belief that his chances were better than her own (an opinion with which he strongly disagreed).

In 2002 she started a master in Data Mining in Pisa and in 2003, after a fortuitous conversation with a midwife during a meeting of a breastfeeding advocacy association, she found herself in a job interview with Eva Buiatti. Eva Buiatti was one of the main Italian epidemiologists and the scientific leader of the then-new-born Agenzia regionale di sanità (ARS), a public health research agency in Florence. Following that interview, Rosa was hired by ARS in a senior position. In her first years at ARS, she developed her skills in epidemiology and big data processing and participated in the development of systems to monitor quality of health care for the elderly.

In 2008, the EU-ADR project started. EU-ADR was a European seminal project in the field of automatic detection of drug safety alerts, and was led by Johan van der Lei and Miriam Sturkenboom, from the Erasmus University Medical Center in Rotterdam, The Netherlands. Rosa was the EU-ADR contributor for ARS together with Giampiero Mazzaglia. In 2009 she participated in the VALORE project, an initiative of the Italian National Agency for the Regional Healthcare Systems (Agenas), in Rome. The project was led by Mariadonata Bellentani and was aimed at assessing the impact of a primary care policy on quality of healthcare for chronic diseases. From the experiences of EU-ADR and VALORE, during a 3-days party for her 40th birthday, she conceived a research project aimed at exploiting Italian administrative databases to monitor quality of health care for chronic diseases. The MATRICE project started in 2011, led by Mariadonata Bellentani, and funded Rosa's new PhD, with Miriam Sturkenboom.

While developing her PhD, she has conducted or participated in multiple activities of ARS, ranging from health services research to pharmacoepidemiology

and data science. Since 2014 ARS has entrusted Rosa with the role of Head of the Pharmacoepidemiology and Medical Informatics Unit.
She lives in Pisa with Maurizio and (now, occasionally) Eloisa and Sara.

ACKNOWLEDGEMENTS

The story that leads to this thesis is punctuated by valuable women, and notable men.

Eva Buiatti, one of the greatest Italian epidemiologists, was a woman with a vision. It took some years for me to fully appreciate how original – and truly visionary, at the time - her initiative of hiring a 'big data' person in a public health agency was. My gratitude for her is only as great as my regret of having so few chances of enjoying her scientific and professional talent, before her premature death. Yet I could feel her gentle touch in many steps I made in the following years.

When Eva asked me, as a favour, to accept to move to a smaller office together with Giampiero Mazzaglia, of course she knew exactly what she was doing. Giampiero is simply brilliant, and the daily conversations with him, over the next years, were probably my most detailed training in the new field of science that I was entering. Notably, Giampiero was at the same time the scientific director of Health Search, and we often compared the characteristics of the administrative data of ARS with the Italian primary care data – from those conversations the core of this book was born. Giampi, you know I owe you!

When I met Mariadonata Bellentani, in Agenas, the common memory of Eva was at the origin of an immediate, complete trust. Her tiny frame masks a steely will, and a stubborn dedication to serve the public health care system. Mariadonata, I can't thank you enough. The chance you gave me in MATRICE was enormous, and I hope you feel that the results are as expected. You have now moved forward in your career, but I am sure that our paths will cross again.

The last gift of Eva, with the complicity of Giampiero, was introducing me in EU-ADR. In this incredibly brave European project I met Miriam – a great scientist, an incredible leader. Miriam, how could I acknowledge all the things that I owe you? They are so many- and some among them are confidential! But there is one episode that I really want to mention. I had just entered in the department, on the eve of an important meeting, planned since months, when I received a call from my brother in Italy: my mother had been admitted

to hospital, she was unconscious and her condition was rapidly worsening. My priorities were immediately shaken, but as many colleagues encouraged me to leave and cancel the meeting, you guessed my feelings and reassured me: "Leave, don't worry: we will hold the meeting online, I will be there for you". On the same night, I was at my mother's bed with my brothers, we discussed with a brilliant clinician, and my mother was saved. The day after, you chaired the online meeting. Miriam, it's a privilege to call you my mentor, and my friend.

That meeting had been organized to ask to Niek Klazinga to join Miriam as my promotor. Niek had been my teacher of Health Services Research at the Erasmus Summer Course, and I had a deep admiration for his clean view of the mechanisms underlying health care systems, rooted in an incomparable experience in international comparison. I was yet to discover his kindness, and to be surprised by the ease with which he mastered the unfamiliar topics I presented him with. Niek, you taught me that science is a social construct, a hard concept for a mathematician. But yes, you are right, and thank you for showing me the 'science' part of that sentence. The very handling of our own conversations – me, you, Miriam, and the fourth of us – was a masterpiece.

The fourth of us is Martijn Schuemie. A guy for whom the time between seeing an interesting problem and designing its solution is just slightly longer than the time needed to develop a prototype and make it work. Martijn, I remember when, in Cape Town, I asked you how a PhD in Rotterdam would have worked. You explained that beyond promoters I would have had a tutor, pointed a finger to yourself and added: for instance, me! That was the moment when I realized that this could indeed happen: you have been the necessary bridge between my formal thinking and this new science. Thank you Martijn – I know it has been hard sometimes.

But this book was born from the interactions with so many others.

First of all, the EU-ADR *task force*: Gianluca Trifirò, Preciosa Coloma, and Paul Avillach. We seamlessly broke new grounds in event harmonization in multi-database studies, and never spent a night sober: thank you so much, guys! EU-ADR was masterfully led by Johan van der Lei, and collected an impressive group of scientists: among them Jan Kors, Erik Mulligen, José Luis Oliveira,

Ron Herings, Annie Fourier-Reglat, Gayo Djallo (Gayo, I will never forget our conversations under a South African sky!), and many others. The project was managed by the magic group of Synapse, who made possible the impossible.

Not less important, the group of Mariadonata in Agenas who conducted MATRICE: Modesta Visca, with her stubborn love for science in healthcare organization, and Giulia Dal Co, a lawyer who quickly learnt to discuss the technical details of data processing. Modesta, Giulia: thanks! Thanks to Mariadonata, VALORE and MATRICE collected a peculiar group of researchers, health care managers, policy makers, who contributed vision, and experience: Andrea Donatini, Gianfranco Damiani, Bruno Federico, Francesco Di Stanislao, Mario Saugo, Carlo Zocchetti, Fulvio Lonati, along with my ARS colleagues. Thanks to Mariagrazia Marvulli and to Mimma Cosentino, and to all data managers of VALORE and MATRICE - Ivan Campa above all.

MATRICE would not have existed without Health Search: thanks to Alessandro Pasqua (you are a friend, Alessandro!), Francesco Lapi, Carmelo Montalbano and, last but not least, Iacopo Cricelli.

MATRICE developed software tools that are deemed to stay. Thanks to the group of CNR led by Raffaele Perego: Emanuele Carlini, Patrizio Dazzi, and thanks to Massimo Coppola, who first saw the feasibility of this project. Thanks to Walter Cazzola and Edoardo Vacchi from the University of Milan. And huge thanks to Fabrizio Carinci and Stefano Gualdi.

Many Italian epidemiologists discussed with me the preliminary results of this thesis. Thanks are due in particular to the BABELE group: Lorenzo Simonato, Marina Maggini, Roberto Raschetti, Nicola Caranci, Valeria Fano, Roberto Gnavi, Cristina Canova. Many inputs came from the group of Carlo Perucci and Marina Davoli in the Dipartimento di Epidemiologia del Lazio, and in particular from Mirko Di Martino, Ursula Kirchmeyer, Valeria Belleudi, Nerina Agabiti, Carla Ancona, Riccardo Didomenicantonio and Nunzia Faustini. The Italian Epidemiology Association is a community of thoughts, thanks in particular to the recent presidents Geppo Costa, Paola Michelozzi, Fabrizio Faggiano.

Over the years, new projects have started, and I had the opportunity to have more scientists join my conversations. Peter Rijnbeek nurtured the Data Derivation Workflow and named the 'component' algorithms: thank you so much Peter! Patrick Ryan, Ingrid Leal, Marius Gheorghe, David Ansell, from EMIF; and Anna Pierini, Rachel Charlton, Helen Dolk, from EUROmediCAT: to all of you I owe an insight. Jeff Brown joined the paper comparing networks by bring the experience of Mini-Sentinel: thanks!

More recently, when the first results of this thesis were published, I was invited to join two impressive groups. In the ADVANCE project I met plenty of great people, and in particular Kaat Bollaerts and Caitlin Dodd, with whom I developed the research plan that starts from here. Thanks, ladies! And in the IMECCHI collaboration I met Hude Quan, whose great papers I had read in the previous years, Bernard Burnand, Bill Ghali, Marie-Annick Le Pogam, Amy Metcalfe, and many others. Thanks to all of you for your encouragement!

To make a thesis you also need your department, and colleagues who create a lively, sometimes challenging environment. So thanks to Sandra and Vera, Coraline and Christel, Nico and Inge, Maria and Ann, Osemeke and René, Marcel and Renske of the cinema nights, Kartini, Alexandra, Gwen, Florentia, Silvana, Benus. Daniel, I am sure we will drink whisky together for many years ahead! Thanks to Peter Moorman for the great gaming nights. Thanks to Desirée, for keeping me on track. Thanks to the secretaries and to the technical staff.

Special thanks to Carmen Ferrajolo and to Gino Picelli, for being so close in a difficult moment.

And now it's time to remind my home base: ARS. Laura Tramonti, Francesco Cipriani, and Paolo Francesconi were the command chain that accepted this proposal, and Fabio Voller later joined in supporting me. Thanks, I could feel Eva in our thoughts when this decision was made. The work environment of ARS is at the same time stimulating and friendly. Thanks to my roommates at the time: Francesco Profili, Valentina Barletta, Matilde Razzanelli, Elena Marchini. Thanks to Andrea Vannucci, Silvia Forni, Manuele Falcone, Francesco Innocenti, Caterina Silvestri, Francesca Collini, Valeria Di Fabrizio, Giacomo

Galletti, Daniela Nuvolone, Pasquale Pepe, Veronica Casotto, Monica da Frè, and Alessandro Barchielli. Thanks to the data and IT groups, led by Marco Santini, Roberto Berni, and Simone Bartolacci. Tiziano Tarli, Daniele Lachi, Giulia Chiarini, Sara Salti, and my dear Claudia Tonon, all make ARS a dynamic institution. Cristina Padovano and Paola Larocca make my travels possible. Thanks to my group: Claudia Bartolini and especially Giuseppe Roberto - Giuseppe, the components wouldn't exist without your energy! Thanks to all the too many that I don't mention, but most of all thanks to the true soul of ARS: Daniela Bachini!

My friends make me. This research project was conceived during one of our parties (thanks Paolo!), and originates from years of free conversation in front of good food and wine, discussing mathematics, science, games, politics, history, music, languages, art. Thanks to all the 'Viarigi' group: you are my community. Thanks to Eloisa senior, to Marco and Omar, to Liviu and Paolo, to Paola and Annalisa, to Ilaria and Gianna, to Giorgia and Fulvio, to Alessandro and Laura, to Francesco 'Cesco' and Elena, to Alessandro 'Salnitro' and Enrico, to Emmanuel and Adrienne. Thanks to all the children in the next generation! Thanks to my dear Javier and to his challenging blog. Thanks to the group of the game of go. Special thanks to Arianna and Rocco, for five months of very successful experience of cohabitation of two families. Most of all, thanks to Zio Minimo: you are home.

My motivation is fed by the memory of some friends that have now gone, but have not been forgotten: Piero, Marco, and Gigi.

For some women the path to professional development starts with the encouragement of their mothers. I am among them. The pain for her loss is soothed by the memory of her pride for my previous academic accomplishments. My family proved strong in the difficult period of her disease: thanks to my brothers Andrea and Alvise, to my sisters-in-law, to my nieces, my aunts, my cousins, and most of all thanks to little Martin, for bringing light again, especially to his grandfather. Thank you dad, for your quiet support.

And finally, thanks to my daughters, Eloisa and Sara. When this work started, you still had an age when mummy not being around for some days was a cause

of discomfort. Yet, you were supportive and loving: thank you. The process was governed by your father. Whenever I was away, he organized for you awesome nights, where all the rules could be broken: stay up till late, eating junk food… everything, to make my absence exciting, rather than painful. Maurizio, you are my true partner in my accomplishments: you have my gratitude, and my love.

LIST OF PUBLICATIONS

Data science, medical informatics and validation studies

Roberto G, Leal I, Sattar N, Loomis AK, Avillach P, Egger P, van Wijngaarden R, Ansell D, Reisberg S, Tammesoo ML, Alavere H, Pasqua A, Pedersen L, Cunningham J, Tramontan L, Mayer MA, Herings R, Coloma P, Lapi F, Sturkenboom2ì M, van der Lei J, Schuemie MJ, Rijnbeek P, **Gini R**. *Identifying cases of type 2 diabetes in heterogeneous data sources: strategy from the EMIF project*. PLoS ONE 2016;11(8):e0160648.

Gini R, Schuemie M, Brown J, Ryan P, Vacchi E, Coppola M, et al. *Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research: A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies.* eGEMs (Generating Evidence & Methods to improve patient outcomes) 2016 ;4(1)

Jonge L de, Garne E, **Gini R**, Jordan SE, Klungsoyr K, Loane M, et al. *Improving Information on Maternal Medication Use by Linking Prescription Data to Congenital Anomaly Registers: A EUROmediCAT Study.* Drug Saf. 2015;1–11.

De Bie S, Coloma PM, Ferrajolo C, Verhamme KMC, Trifirò G, Schuemie MJ, Straus SMJM, **Gini R**, et al. *The role of electronic healthcare record databases in paediatric drug safety surveillance: a retrospective cohort study*. Br J Clin Pharmacol. 2015 ;80(2):304-14.

Patadia VK, Coloma P, Schuemie MJ, Herings R, **Gini R**, Mazzaglia G, et al. *Using real-world healthcare data for pharmacovigilance signal detection - the experience of the EU-ADR project*. Expert Rev Clin Pharmacol. 2015;8(1):95–102.

Valkhoff VE, Coloma PM, Masclee GMC, **Gini R**, Innocenti F, Lapi F, et al. *Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk.* Journal of clinical epidemiology 2014;67(8):921–31.

Charlton RA, Neville AJ, Jordan S, Pierini A, Damase-Michel C, Klungsøyr K, Nybo Andersen AM, Hansen AV, **Gini R**, et al. *Healthcare databases in Europe for studying medicine use and safety during pregnancy*. Pharmacoepidemiology and Drug Safety 2014;23(6):586-94.

Gini R, Schuemie MJ, Francesconi P, Lapi F, Cricelli I, Pasqua A, et al. *Can Italian healthcare administrative databases be used to compare regions with respect to compliance with standards of care for chronic diseases?* PLoS ONE 01/2014; 9(5):e95419.

Gini R, Francesconi P, Mazzaglia G, Cricelli I, Pasqua A, Gallina P, et al. *Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey.* BMC Public Health 2013; 13(1):15.

Avillach P, Coloma PM, **Gini R**, Schuemie M, Mougin F, Dufour J-C, et al. *Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project.* Journal of the American Medical Informatics Association 2013;20(1):184-92.

Caranci N, Fano V, **Gini R**, Maggini M, Raschetti R, Simonato L. *[A laboratory to overcome the babel of the electronic health archives].* Epidemiologia e prevenzione 2012; 36(5):234-5.

Schuemie MJ, **Gini R**, Coloma PM, Straatman H, Herings RMC, Pedersen L, et al. *Replication of the OMOP Experiment in Europe: Evaluating Methods for Risk Identification in Electronic Health Record Databases*. Drug Safety. 2013;36(S1):159–69.

Schuemie MJ, Coloma PM, Straatman H, Herings RMC, Trifirò G, Matthews JN, Prieto-Merino D, Molokhia M, Pedersen L, **Gini R**, Innocenti F, Mazzaglia G, Picelli G, Scotti L, van der Lei J, Sturkenboom M. *Using Electronic Health Care Records for Drug Safety Signal Detection: A Comparative Evaluation of Statistical Methods.* Medical care 2012; 50(10):890-897.

Coloma PM, Trifirò G, Schuemie MJ, **Gini R**, Herings R, Hippisley-Cox J, et al. *Electronic healthcare databases for active drug safety surveillance: is there enough leverage?*. Pharmacoepidemiology and Drug Safety 02/2012; 21(6):611-21.

Coloma PM, Schuemie MJ, Trifirò G, **Gini R**, Herings R, Hippisley-Cox J, et al. *Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project.* Pharmacoepidemiology and Drug Safety 2011; 20(1):1-11.

Avillach P, Joubert M, Thiessard F, Trifirò G, Dufour J-C, Pariente A, Mougin F, Polimeni G, Catania MA, Giaquinto C, Mazzaglia G, Fornari C, Herings R, **Gini R**, Hippisley-Cox J, Molokhia M, Pedersen L, Fourrier-Réglat A, Sturkenboom M, Fieschi M. *Design and evaluation of a semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European EU-ADR project.* Studies in health technology and informatics 2010; 160(Pt 2):1085-9.

Avillach P, Mougin F, Joubert M, Thiessard F, Pariente A, Dufour J-C, Trifirò G, Polimeni G, Catania MA, Giaquinto C, Mazzaglia G, Baio G, Herings G, **Gini R**, Hippisley-Cox J, Molokhia M, Pedersen L, Fourrier-Réglat A, Sturkenboom M, Fieschi M. *A semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European EU-ADR project.* Stud Health Technol Inform. 2009;150:190–4.

Gini R. *Stata tip 56: Writing parameterized text files*. Stata Journal 2008; 8(1):134-136.

Francesconi P, **Gini R**, Roti L, Bartolacci S, Corsi A, Buiatti E.*The Tuscany experimental registry for Alzheimer's disease and other dementias: how many demented people does it capture?* Aging clinical and experimental research 2007; 19(5):390-3.

Gini R, Pasquini J. *Automatic generation of documents*. Stata Journal 2006; 6(1):22-39.

Trifirò G, Patadia V, Schuemie MJ, Coloma PM, **Gini R**, Herings R, et al. *EU-ADR healthcare database network vs. spontaneous reporting system database: preliminary comparison of signal detection.* Studies in health technology and informatics 2011; 166:25-30.

Health services research , policy evaluation and occupational epidemiology

Buja A, Solinas G, Visca M, Federico B, **Gini R**, Baldo V, et al. *Prevalence of Heart Failure and Adherence to Process Indicators: Which Socio-Demographic Determinants are Involved?* International Journal of Environmental Research and Public Health. 2016;13(2):238.

Buja A, **Gini R**, Visca M, Damiani G, Federico B, Donato D, et al. *Need and disparities in primary care management of patients with diabetes*. BMC Endocrine Disorders. 2014;14(1):56.

Buja A, Damiani G, **Gini R**, Visca M, Federico B, Donato D. *Systematic age-related differences in chronic disease management in a population-based cohort study: a new paradigm of primary care is required.* PLoS ONE 2014; 9(3):e91340.

Visca M, Donatini A, **Gini R**, Federico B, Damiani G, Francesconi P, et al. *Group versus single handed primary care: A performance evaluation of the care delivered to chronic patients by Italian GPs.* Health Policy 2013 113(1-2):188-98.

Buja A, **Gini R**, Visca M, Damiani G, Federico B, Francesconi P, et al. *Prevalence of chronic diseases by immigrant status and disparities in chronic disease management in immigrants: a population-based cohort study, Valore Project.* BMC Public Health 2013; 13(1):504.

Costa G, Michelozzi P, Ancona C, Bertozzi N, Caranci N, Fano V, **Gini R**, Gnavi R, Zocchetti C. *[Health and economic crisis in Italy].* Epidemiologia e prevenzione 2012; 36(5):227-8.

Levi M, Baldasseroni A, Biffino M, **Gini R**, Romeo G, Paggiaro P, Talini D. *Burden of disease of occupational asthma in Tuscany: Methodological aspects and preliminary results*. Giornale italiano di medicina del lavoro ed ergonomia 2012; 34(3):240-244.

Francesconi P, **Gini R**, Maciocco G, Damiani G. [*Primary care and chronic diseases: geographical differences in avoidable hospitalization].* Epidemiologia e prevenzione 2010;35(5-6 Suppl 2):128-9.

Carinci F, Roti L, Francesconi P, **Gini R**, Tediosi F, Di Iorio T, Bartolacci S, Buiatti E. *The impact of different rehabilitation strategies after major events in the elderly: the case of stroke and hip fracture in the Tuscany region.* BMC Health Services Research 2007; 7:95.

Gini R, Capon A, Roti L, Mastromattei A, Buiatti E. [*Femur fractures among elderly in Lazio and Tuscany regions from 1999 to 2003].* Epidemiologia e prevenzione 2007; 31(4):197-203.

Pharmacoepidemiology

Charlton RA, Klungsøyr K, Neville AJ, Jordan S, Pierini A, Berg LTW de J den, Bos HJ, Puccini A, Engeland A, **Gini R**, et al. *Prescribing of Antidiabetic Medicines before, during and after Pregnancy: A Study in Seven European Regions.* PLoS ONE. 2016;11(5):e0155737.

Marcianò I, Ingrasciotta Y, Giorgianni F, Bolcato J, Chinellato A, Pirolo R, Di Giorgio A, Manna S, Ientile V, **Gini R** , et al. *How did the Introduction of Biosimilar Filgrastim Influence the Prescribing Pattern of Granulocyte Colony-Stimulating Factors? Results from a Multicentre, Population-Based Study, from Five Italian Centres in the Years 2009-2014*. BioDrugs. 2016;30(4):295-306.

Charlton RA, Pierini A, Klungsøyr K, Neville AJ, Jordan S, Berg LTW de J den, Thayer D, Bos HJ, Puccini A, Hansen AV, **Gini R** , et al. *Asthma medication prescribing before, during and after pregnancy: a study in seven European regions*. BMJ Open. 2016;6(1):e009237.

Coloma PM, de Ridder M, Bezemer I, Herings RMC, **Gini R**, Pecchioli S, et al. *Risk of cardiac valvulopathy with use of bisphosphonates: a population-based, multi-country case-control study*. Osteoporos Int. 2015 27(5):1857-67.

Bezemer ID, Verhamme KMC, **Gini R**, Mosseveld M, Rijnbeek PR, Trifirò G, et al. *Use of oral contraceptives in three European countries: a population-based multi-database study.* The European Journal of Contraception & Reproductive Health Care. 2015;21(1):81-7

Pacurariu AC, Straus SM, Trifirò G, Schuemie MJ, **Gini R**, Herings R, et al. *Useful Interplay Between Spontaneous ADR Reports and Electronic Healthcare Records in Signal Detection*. Drug Saf. 2015;38(12):1201-10.

Charlton R, Garne E, Wang H, Klungsøyr K, Jordan S, Neville A, Pierini A, Hansen A, Engeland A, **Gini R**, et al. *Antiepileptic drug prescribing before, during and after pregnancy: a study in seven European regions*. Pharmacoepidemiol Drug Saf. 2015.

Ingrasciotta Y, Giorgianni F, Bolcato J, Chinellato A, Pirolo R, Tari DU, Troncone C, Fontana A, Ientile V, **Gini R**, et al. *How Much Are Biosimilars Used in Clinical Practice? A Retrospective Italian Population-Based Study of Erythropoiesis-Stimulating Agents in the Years 2009-2013*. BioDrugs. 2015.

Charlton R, Jordan S, Pierini A, Garne E, Neville A, Hansen A, **Gini R**, et al. *Selective serotonin reuptake inhibitor prescribing before, during and after pregnancy: a population-based study in six European regions.* BJOG 2014 ;24(11):1144-54.

Masclee GM, Valkhoff VE, Coloma PM, de Ridder M, Romio S, Schuemie MJ, Herings R, **Gini R**, Mazzaglia G, Picelli G, Scotti L, Pedersen L, Kuipers EJ, van der Lei J, Sturkenboom M. *Risk for Upper Gastrointestinal Bleeding from Different Drug Combinations.* Gastroenterology 2014;147(4):784-792.

Ferrajolo C, Coloma PM, Verhamme KMC, Schuemie MJ, de Bie S, **Gini R**, et al. *Signal detection of potentially drug-induced acute liver injury in children using a multi-country healthcare database network*. Drug Safety 2014; 37(2):99-108.

Coloma PM, Schuemie MJ, Trifirò G, Furlong L, van Mulligen E, Bauer-Mehren A, Avillach P, Kors J, Sanz F, Mestres J, Oliveira JL, Boyer S, Helgee EA, Molokhia M, Matthews J, Prieto-Merino D, **Gini R**, Herings R, Mazzaglia G,

Picelli G, Scotti L, Pedersen L, van der Lei J, Sturkenboom M. *Drug-Induced Acute Myocardial Infarction: Identifying 'Prime Suspects' from Electronic Healthcare Records-Based Surveillance System.* PLoS ONE 2013; 8(9).

Software

TheMatrix: the tool for easy observational health data management from CSV files. 2015. http://thematrix.isti.cnr.it/

FUNNELCOMPAR: Stata module to perform funnel plot for institutional comparison. 2009. https://ideas.repec.org/c/boc/bocode/s457078.html

REWRITE: Stata module to rewrite text files from disk performing macro substitutions. 2009. https://ideas.repec.org/c/boc/bocode/s457066.html

Mathematics

Gini R, Ornea L, Parton M, Piccinni P. *Reduction of Vaisman structures in complex and quaternionic geometry*. Journal of Geometry and Physics 2006;56 (12):2501-2522

Gini R, Ornea L, Parton M. *Locally conformal Kähler reduction*. Krelle's Journal 2005;(581):1-21.

Funar L, **Gini R**. *The Graded Cobordism Group of Codimension-One Immersions*. Geometric and Functional Analysis 2002; 12(6):1235-1264.

Gini R. *Cobordism of immersions of surfaces in non-orientable 3-manifolds*. manuscripta mathematica 2000; 104(1):49–69.

PhD PORTFOLIO

Name	Rosa Gini
Erasmus MC Department	Medical Informatics
Research School	Netherlands Institute for Health Sciences
PhD period	February 2011 to October 2016
Promotors	Prof.dr. M.C.J.M. Sturkenboom
	Prof.dr. N.S. Klazinga

1. Training

Research skills

21st Erasmus Summer Programme. Rotterdam, August 15 – September 2, 2011

Oral presentations

2015	*Un'infrastruttura informatica a supporto di studi epidemiologici multicentrici su sistemi informativi sanitari: il software open source TheMatrix.* XXXIX Conference of the Italian Epidemiology Association. Milan, October 2015.
	Identifying chronic conditions from data sources with incomplete diagnostic information: the case of Italian administrative databases. RECIF Meeting. Aarhus, Denmark. September 2015.
2014	*Validazione di algoritmi per individuare diabete, ipertensione e cardiopatia ischemica dai database amministrativi italiani: lo studio MATRICE*. XXXVIII Conference of the Italian Epidemiology Association. Naples, November 2014.
	Validating case definitions of chronic diseases in Italian regional health service databases: results from the MATRICE Project. Doing Research in Healthcare with Administrative Databases. Milan, June 2014
2013	*Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies For Data Extraction And Management.* 29th International Conference on Pharmacoepidemiology and Therapeutic Risk Management. Montreal, August 2013.
2011	*Variabilità geografi ca di indicatori di processo nella cura di diabete, insufficienza cardiaca e cardiopatia ischemica: confronto tra stime ottenute da dati amministrativi correnti e stime ottenute dai dati clinici della medicina generale nel progetto VALORE*. XXXV Conference of the Italian Epidemiology Association. Turin, November 2011.
2010	*Harmonising definitions of adverse events among 8 European healthcare databases participating in the EU-ADR Project.* EUROEPI 2010 Epidemiology and Public Health in an Evolving Europe. November, 2010, Florence, Italy.

Posters

2016	*Reliably estimating adherence to therapy in chronic patients from Italian Administrative Databases*. 32nd International Conference on Pharmacoepidemiology and Therapeutic Risk Management. Dublin, August 2016.
	Standardizing diversity of event definitions in federated networks of heterogeneous health data sources: the "component algorithm strategy". 32nd International Conference on Pharmacoepidemiology and Therapeutic Risk Management. Dublin, August 2016.
2015	*Identifying chronic conditions from data sources with incomplete diagnostic information: the case of Italian administrative databases*. 31st International Conference on Pharmacoepidemiology and Therapeutic Risk Management. Boston, August 2015
	Automatic identification of stages of type 2 diabetes, hypertension, ischaemic heart disease and heart failure from Italian General Practioners' electronic medical records: a validation study. 31st International Conference on Pharmacoepidemiology and Therapeutic Risk Management. Boston, August 2015
	Identifying cases of type 2 diabetes fromheterogeneous data sources: strategy from the EMIF Project. 31st International Conference on Pharmacoepidemiology and Therapeutic Risk Management. Boston, August 2015

Workshops

2015	*Assistenza primaria: monitoraggio della qualità e valutazione d'impatto nei progetti Agenas MATRICE - MEDINA – LUNA*. Final conference of the MATRICE Project. Rome, May 2015.
2012	*Frameworks for Data Extraction and Management from Electronic Healthcare Databases for Multi-Center Epidemiologic Studies: a Comparison among EU-ADR, MATRICE, and OMOP Strategies.* Medical Informatics in Europe 2012. Pisa, August 2012.
2010	*Design, development and validation of a computerised system that exploits data from electronic health records and biomedical information for the early detection of adverse drug reactions. The EU-ADR project: Preliminary Results.* 13th International Congress on Medical Informatics (Medinfo 2010). Cape Town, South Africa, September 2010.

Other seminars

2016-today	Journal Club of the Pharmacoepidemiology and Medical Informatics Unit, ARS, Florence
2011-today	Research Seminars in Pharmacoepidemiology, Rotterdam Research Colloquia in Medical Informatics, Rotterdam
2015	*Validazione di algoritmi per individuare diabete, ipertensione e cardiopatia ischemica dai database amministrativi italiani: lo studio MATRICE.* Department of Health Statistics, Università Bicocca. Milan, July 2015.
	Identify diseases from data sources where information on diagnosis is incomplete: experiences from Europe and beyond. Centre hospitalier universitaire vaudois. Lausanne, Switzerland, June 2015

Other

2014-today	Member of the International Methodology Consortium for Coded Health Information
2011-today	Reviewer for Drug Safety, International Journal of Quality of Care, BMC Public Health, Health Policy, PLoS ONE, BMJ Open, ICPE
2011-today	Member of the International Society for Pharmacoepidemiology
2007-today	Member of the Italian Epidemiology Association
2010-2012	Member of the Italian Epidemiology Association Secretariat

2. Teaching

2014	Uso delle fonti di dati sanitari correnti per finalità epidemiologiche. Istituto Superiore di Sanità. Rome, October 22-24.
2013	Statistica per economia, scienze sociali, epidemiologia clinica e sanità pubblica. Rome, June 3-7