BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (http://bmjopen.bmj.com).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

Electronic healthcare databases in Europe: characterisation and assessment of usefulness for benefit-risk evaluation of medicines

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-023090
Article Type:	Research
Date Submitted by the Author:	20-Mar-2018
Complete List of Authors:	Pacurariu, Alexandra; European Medicines Agency, Pharmacovigilance and Epidemiology Department; Dutch Medicines Evaluation Board, Plueschke, Kelly; Pharmacovigilance and Epidemiology McGettigan, Patricia; Queen Mary University of London, William Harvey Research Institute; European Medicines Agency Morales, Daniel; European Medicines Agency (EMA), Surveillance & Epidemiology Service; University of Dundee, Division of Population Health Sciences Slattery, Jim; European Medicines Agency, Surveillance & Epidemiology Service Vogl, Dagmar; European Medicines Agency Goedecke, Thomas; European Medicines Agency Kurz, Xavier; European Medicines Agency, Cave, Alison; European Medicines Agency, Pharmacovigilance and Epidemiology Department
Keywords:	electronic healthcare databases, post-authorisation studies, regulatory science, benefit-risk evaluation, real-world data, EPIDEMIOLOGY

SCHOLARONE™
Manuscripts

12th March 2018

Electronic healthcare databases in Europe: characterisation and assessment of usefulness for benefit-risk evaluation of medicines

Alexandra Pacurariu*[1,2], Kelly Plueschke[1], Patricia McGettigan[1,3], Daniel Morales[1,4], Jim Slattery[1], Dagmar Vogl[1], Thomas Goedecke[1], Xavier Kurz[1], Alison Cave[1]

1 Surveillance & Epidemiology Service, European Medicines Agency (EMA), 30 Churchill Place, London E14 5EU, United Kingdom

2 Dutch Medicines Evaluation Board, 3531 AH Utrecht, the Netherlands

3 William Harvey Research Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK

4 Division of Population Health Sciences, University of Dundee, Dundee, United Kingdom

* Corresponding author: Alexandra Pacurariu alexandra.pacurariu@ema.europa.eu, +44(0)2036608559

Word count: 3662

Number of figures and tables:

3 figures and 2 tables in the manuscript

Key words:

Electronic healthcare databases; post-authorisation studies; regulatory science; benefit-risk evaluation; real-world data; epidemiology.

For peer review only

Abstract

Objective

Electronic healthcare databases (EHDs) are important for drug development and safety evaluation but they are heterogeneous and fragmented in Europe. In this paper, we describe the process used to identify, describe and evaluate the usefulness of EHDs databases to support regulatory decisions.

Methods

EHDs were identified from publicly available information such as the ENCePP resources database, textbooks, and web-based searches. Databases were selected using criteria such as accessibility, longitudinal dimension, recording of exposure and outcomes and generalizability. Extracted data was verified with the database owners, when possible.

Results

A total of 34 EHDs were identified. The most represented regions were Northern, Central and Western Europe. The most frequent type of data source was electronic medical records (44.1%) and record linkage systems (29.4%). The median number of patients registered in the 34 data sources was 5 million (range 0.07-15 million) while the median calendar time covered by a database was 18.5 years.

Conclusion

There is significant heterogeneity between the European EHDs in terms of structure, content and duration. There is a scarcity of data sources in some areas, particularly Eastern Europe, and a lack of information from the secondary care setting.

For peer review only

Strengths and limitations of this study

- Incomplete data capture might - although unlikely, cannot be completely excluded, especially for very small and new databases
- Data extraction was mostly based on publicly available information and there might be inconsistencies or outdated data
- Validation of the data source was only evaluated indirectly through the number of validation studies reported by the database owners

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1. Introduction

The European Union (EU) medicines regulatory network has responsibility for protecting patients by ensuring continuous evaluation of the safety of authorised medicines. At the core of such review is the scientific assessment of all available evidence including relevant information from the literature, results from non-clinical studies, randomised clinical trials (RCTs), observational studies, spontaneous reports and other available research. A way to collect more information about a medicine's safety post-marketing is by means of post-authorisation safety studies (PASS).(1) PASS may be imposed on a marketing authorisation holder by a regulatory authority, or voluntarily proposed by the company in the medicinal product's risk management plan to address a safety concern or evaluate the effectiveness of risk minimisation measures aimed at reducing the occurrence or severity of an adverse reaction.(2,3)

Secondary use of routinely collected data from electronic healthcare databases (EHDs) is often used in such studies because it is usually faster and cheaper than primary data collection.

A review of pharmaceutical industry-sponsored studies evaluating the effectiveness of risk minimisation measures submitted to the European Medicines Agency (EMA) for cardiovascular, endocrinology or metabolic drugs authorised between 1995 and 2015 found that EHDs were used in 53% and 31% of studies evaluating routine and additional risk minimisation measures. (4) A second review of 189 PASS assessed by the EMA between 2012 and 2015 and registered in the European Union electronic Register of Post-Authorisation Studies (EU PAS Register®) showed that secondary use of routinely collected data occurred in 33.3% of instances (results based on a sub-sample of 57 studies with the protocol available in the EU PAS Register®).(5) Among the 19 (33%) PASS using secondary data collection, 58% leveraged electronic health records (EHRs). A third review of a different set of studies registered in the EU PAS Register® as of December 2016 found that 117 studies (37%), used an existing claims or electronic medical records database. (6) A fourth review, this time evaluating studies which measured the impact of regulatory interventions, found that claims databases were used in 45% of studies, while EHRs were used in 22%, the latter being the most utilised type of data sources for such studies.(7)

The frequent use of EHDs in observational studies was also reported in a wider context in a review of the abstracts of presentations made at the International Conference for Pharmacoepidemiology: 53% (in 2000) and 51% (in 2005) of submitted EU pharmacoepidemiological studies were conducted using automated general practice, pharmacy or claims data.(8) The fact that between 30%-50% of

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

observational post-authorisation studies use EHDs as their main data source reflects the importance of these data sources to support regulatory decision making. (1,9)

This use of EHDs starts also to be reflected in pre-authorisation research. However, our experience is that there is currently limited use of real world data pre-authorisation, while its use is mostly focused on historical control data or understanding the natural history of the disease.

As regulatory decisions based on EHDs may have a high impact on public health, the quality of the information contained in the databases and the validity and reproducibility of the derived results require close attention, especially when combining data from several data sources or when the original data is transformed before analysis. (10–12) It has been emphasised that the same level of scientific rigour should be employed irrespective of the study design and data source to be used, and that the strengths and weaknesses of each data source should be considered. (13) The speed at which the results could be generated is an additional valid consideration, particularly for regulatory purposes.(9,14,15) By considering the characteristics of the data sources and the research objectives to be addressed, the investigators should be able to choose the most appropriate resource(s) to address the question at hand. However, while some authors provide a detailed description of the databases used in their study, (16–19) reporting in the published literature is often incomplete and rarely a justification for the choice in the context of alternative data sources provided. (18) The International Society of Pharmacoepidemiology (ISPE) has developed guidelines to support the selection and use of data sources for observational research by highlighting potential limitations of databases and recommending testing procedures. The guidelines also provide a checklist covering six areas: database selection, use of multiple data resources, extraction and analysis of the study population, privacy and security, quality and validation procedures and documentation.(20) The availability of an inventory of European databases describing the main characteristics, conditions of access and validation performed would support investigators to identify databases suitable for their research question. Moreover, knowledge of the characteristics of the data sources used in a post-authorisation study would enhance regulators' confidence in the evidence derived from such data and ultimately in the usefulness of the study in the decision-making process.(21,22)

The main objective of this study is to provide an inventory of EHDs and describe their key characteristics with the aim to support stakeholders in their choice to conduct a post authorisation study using such source of information.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

2. Methods

2.1. Identification of EHDs

As a first step, we identified existing EHDs in Europe by screening the following sources: the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) resources database, (21) web-based search engines dedicated to healthcare databases (22), textbooks on clinical pharmacoepidemiology (23,24), publicly available inventories created for European Commission funded research projects and databases used in publicly funded research such as EMA funded post-authorisation studies. The different population registries within the same country were considered as a single national EHD if they could be linked using a unique identification number (e.g. Nordic countries, Scotland).

As a second step, data sources were included based on the following criteria: the data is available to regulatory authorities or to third-parties for research purposes; the database contains information on both drug exposure and health outcomes and is not disease or product specific; there is longitudinal data capture. Prescription-only databases were excluded because they cannot be used alone for aetiological studies. Product or disease specific registries were considered out of scope as they create cohorts of patients whose entry is defined either by exposure to a product or by occurrence of a disease or health outcome.(25)

Product specific registries are frequently used for the benefit-risk monitoring of specific products, they rarely cover a wide range of medicines and health conditions and the same registry cannot be used to monitor medicinal products of different therapeutic classes. Databases where the data collection ceased and the historical data is not accessible were excluded.

2.2. Data extraction and classification

For each database publicly available information (on the databases' websites or in publications) was supplemented by contacting data source owners in writing. A total of 82% database owners responded. Teleconferences with seven database owners were conducted to clarify some of the information provided.

The information was extracted by six EMA reviewers (AP, KP, PMG, DM, AC, JS) and the entries for each database were cross-checked for consistency by a second reviewer. Uncertainties about the classification of any variable were resolved through discussion.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

The data sources were classified in three categories according their structure, purpose, and type of data: electronic medical records, claims databases or healthcare record linkage systems (e.g. several databases are linked to form a complete database). The size of the data source was quantified by the cumulative number of patients included (both total and active patients) and number of years since the initiation of data collection in the database.

Data collected in the following categories was also recorded: demographic information (age and gender of each individual), information on prescribed or dispensed medicines (including name, dose, duration, route of administration and therapeutic indication), immunisations, diagnosis data and referrals for laboratory investigations, imaging, and other procedures. Information of laboratory tests results was not collected.

2.3. Availability of validation studies

We searched the literature for validation studies for the databases included and database owners were asked to report validation studies they were aware of. For the purpose of this study, a validation study was defined as any study published in a peer-reviewed journal that aimed to validate the information available on an outcome or exposure in comparison with gold standard information, usually the patients' original health records as reviewed by a medical professional or valid information from another database capturing the same information for a different purpose. For example, a study in 2012 compared cancer recording in a general practitioners' database, hospital records and cancer registries and found considerable discrepancies in cancer recording between these different data sources.(26) The number of validation studies performed was used as an indicator of the overall validity of the database but there was no quality assessment of each individual study.

2.4. Accessibility

The accessibility of databases for research purposes was classified in four categories: no access, indirect access through the database owner or a third party, direct access restricted to specific datasets and direct access to the full dataset.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

2.5. Coding of database characteristics for usefulness for medicines benefit-risk evaluation

In order to assist in the selection of databases (rather than evaluating the quality of each database), a coding process was undertaken to identify the data sources considered to provide sufficient information to contribute to regulatory questions on the benefit-risk evaluation of medicines. For this purpose, the following domains were included in the coding process: extent of data capture of study variables, size of data source, quality and validity of information, data format, accessibility and potential for linkage (Figure 1). The coding system was in most cases binary: 0 if information was absent and 1 if it was present. An exception was the accessibility variable for which the following categories were assigned: 0 - no access; 1 - indirect access through database owner or third party; 2 - direct access to specific data sources; 3 - direct access to full data source. The validity variable was scored: 0 - no validation studies present; 1 - less than 5 studies; 2 - more than 5 studies. The degree of completion for a specific variable was not recorded, only if the variable was present in database structure or not.

2.6. Common Data Model (CDM)

A common data model provides a common representation and architecture of the data across multiple databases, thus enabling the standardisation of administrative and clinical information and allowing the use of common analytical tools. (27)

[Insert Figure 1. Scoring of the usefulness of electronic healthcare databases available in Europe for the benefit-risk evaluation of medicines.]

The initial inventory with the description of databases was reviewed by the ENCePP Working Group "Data Sources". (9) Following amendments, the final inventory was endorsed by the ENCePP Steering Group.

3. Results

3.1. General overview

The initial search generated a list of 77 potential data sources. After merging the national registries into a single entry where applicable the exclusion criteria were applied to 48 databases and 34 were retained in the final inventory (Figure 2). A summary description of these 34 databases is presented in Table 1 and the complete information is provided in the *Supplementary material.*

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

[Insert Figure 2. Flowchart of database selection]

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 1. List of data sources retained in the final inventory (by year)

Data source name	Country	Type	Type of care	Start date
Finnish National registries	Finland	Record linkage system	Mixed	1964
Swedish National registries	Sweden	Record linkage system	Mixed	1970
Danish National and regional registries	Denmark	Record linkage system	Mixed	1977
The electronic Data Research and Innovation Service	Scotland	Record linkage system	Mixed	1981
Clinical Practice Research Datalink	United Kingdom	Electronic medical records	Primary care	1987
QResearch	United Kingdom	Electronic medical records	Primary care	1989
Information System of Parc de Salut del Mar	Spain	Electronic medical records	Primary care	1990
Medicines Monitoring Unit Scotland	Scotland	Record linkage system	Mixed	1990
Pharmo Database Network	Netherlands	Record linkage system	Mixed	1990
QuintilesIMS Disease Analyser	Germany	Electronic medical records	Mixed	1992
Integrated Primary Care Information Database	Netherlands	Electronic medical records	Primary care	1995
Agencia Regionale di Sanita Tuscany database	Italy	Claims	Secondary care	1996
Norwegian Registries	Norway	Record linkage system	Mixed	1997
QuintilesIMS Disease Analyser	France	Electronic medical records	Primary care	1997
Region Emilia Romagna Database	Italy	Claims	Secondary care	1997
Hospital Information System -Lazio	Italy	Claims	Secondary care	1998
Icelandic Registries	Iceland	Record linkage system	Mixed	1998
Pedianet Database	Italy	Electronic medical records	Primary care	1998
Securite Sociale de l'Assurance Maladie	France	Claims	Mixed	1999
Lombardia Health Database	Italy	Electronic medical records	Secondary care	2000
QuintilesIMS LPD Health Search Database Longitudinal	Italy	Electronic medical records	Primary care	2000
QuintilesIMS LifeLink: Hospital Disease Database - Belgium	Belgium	Electronic medical records	Secondary care	2001
QuintilesIMS LPD Health Search Database Longitudinal	France	Electronic medical records	Mixed	2001
Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria	Spain	Electronic medical records	Primary care	2002
Caserta Database	Italy	Claims	Primary care	2002
The Health Improvement Network	United Kingdom	Electronic medical records	Primary care	2002
German Pharmacoepidemiological Research Database	Germany	Claims	Mixed	2004
Echantillon Généraliste de Bénéficiaires	France	Claims	Mixed	2006
QuintilesIMS LPD Health Search Database Longitudinal	Spain	Electronic medical records	Mixed	2006
The Information System for the Development of Research in Primary Care	Spain	Electronic medical records	Primary care	2006
VEKTIS	Netherlan	Claims	Mixed	2006

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	ds			
Secure Anonymised Information Linkage	Wales	Record linkage system	Mixed	2007
National Health Fund	Poland	Claims	Mixed	2008
Hospital Treatment Insights	United Kingdom	Record linkage system	Secondary care	2010

The most frequent type of data source identified were electronic medical records (n=15, 44.1%) followed by record linkage system (n=10, 29.4%) and claims databases (n=9, 26.5%). In terms of type of care covered, mixed care settings (primary and secondary care) were most common (n=17, 50%), followed by primary care databases (n=11, 32.3%) (see Table 2). The median number of patients followed cumulatively across the 34 data sources was 5 million (range 0.07-15 million).

Table 2. Distribution of data sources type and type of care covered

Type of data source	Primary care	Secondary care	Mixed
Claims	1 (2.9%)	3 (8.82%)	5 (14.7%)
Electronic medical records	10 (29.4%)	2 (5.9%)	3 (8.8%)
Record linkage system	0 (0%)	1 (2.9%)	9 (26.5%)

Patient age and gender were recorded in all data sources while paediatric patients were included in 32 databases (94%). The median year for database start was 1998, with the oldest database established in 1964 (the Finnish Hospital Discharge Register). The median calendar time covered by a database was 18.5 years (range 7-53 years). In terms of geographical coverage, 17% of databases collect data from Norway, 14% from Finland and 10% from Denmark and Italy (see Figure 3).

[Insert Figure 3. Number of data sources across Europe and duration of data collection
Boxplots indicate the median (horizontal black line) while the margins of the boxplot represent the interquartile range, the vertical lines indicate the minimum and maximum value. The number of databases per country are also indicated above the boxplots.]

3.2. Information captured

By defin 1 , all the databased retained in the final inventory contained information about drug exposure (either prescribed or dispensed). However, the completeness of information was variable: 28 (82.3%) databases had information about prescribed dose and duration of treatment (either directly recorded or inferred from other collected variables); 14 (41.1%) had information about route of

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

administration; 20 (58.8%) of the databases recorded the therapeutic indication associated with the prescription (either directly recorded or inferred from other database elements). Over-the-counter drugs were not captured in any of the databases while vaccinations were captured in 13 databases (38%). Data on hospital in-patient administered drugs were rarely captured (5.8%).

All databases had information about medical events (diagnosis) as a prerequisite for inclusion in our inventory. Referrals for laboratory investigations were captured in 19 (56%) and referrals for imaging or other diagnostic procedures were captured in 16 (47%) databases.

3.3. Validation status

For one third of the databases no published validation study could be identified; less than 5 validation studies were found for one third of databases whereas one third had more than 5 validation studies.

3.4. Accessibility and potential for linkage

Only one database was excluded due to the fact that no third party access was allowed. From the remaining ones, 32% of EHDs offered indirect access to the database for third parties, 21% provide direct access to specific datasets and 24% offered direct access to the full datasets. The level of access was unknown for 23% of EHDs. In terms of linkage, 68% of the databases could be linked through a unique personal identification number (PIN) to other databases containing additional healthcare-related information including cause of death registries, hospital data, prescription databases and cancer registries. The Nordic registries are a good example of extensive linkage among different national registries by usage of PIN. Other forms of linkage were sometimes used. For example in order to avoid the use of PIN to preserve anonymity, the PHARMO network uses probabilistic linkage based on patient birth date, gender and general practitioner code. The linkage of a parent with their child ('parent-child linkage'), which is useful for studies investigating pregnancy exposures and effect on offspring, was available in 7 data sources (21%).

3.5. Transformation of the database to a CDM

Four (11.7%) databases were already transformed in a CDM, while other four were in process of being transformed to a CDM (QuintilesIMS Disease Analyser France and Germany, Spanish Information System for the Development of Research in Primary Care, Agenzia Regionale di Sanità Tuscany database, Clinical Practice Research Datalink, Integrated Primary Care Information Database and The

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Health Improvement Network). Seven out of eight databases used the Observational Medical Outcomes Partnership CDM. (27)

4. Discussion

A total of 34 European EHDs with potential to be used in the regulatory environment were identified in this review. The regions most frequently represented were Northern, Central and Western Europe, with a scarcity of data sources in Eastern Europe. The most common data sources assessed were electronic medical records with a mix of primary and secondary care coverage. Most of the databases contain out-patient prescribing while inpatient drug administration is captured very rarely. The median number of patients registered within the 34 data sources was 5 million and the median calendar time covered by a database was 18.5 years. In terms of accessibility, 24% of databases offered direct access to the full data source, with the rest having a somewhat more limited access. Over 60% of databases had validation studies published.

There are a few similar reviews of EHDs available in Europe (8,28) and as far as we are aware this is the first review with a regulator perspective. An analysis of the characteristics of post-authorisation studies requested by regulators showed that 47% of studies involved secondary use of data emphasizing the important role of secondary data in regulatory setting.

Given that it is well known that different national practices driven by local guidelines and clinical practice can generate significant heterogeneity in how healthcare is delivered and recorded (29), it is important that regulators have access to data from as broad a geographical spread as possible. Thus there is a clear need for the development of data sources in EU member states which currently either have no data sources or are poorly represented.

With regards to the data recorded in the databases, a few limitations are observed. Firstly, the limited capture of in-patient prescribing poses a problem for regulators and investigators since many newly approved drugs are specialist drugs, used exclusively in secondary care (30). Secondly, some disease-specific variables (e.g., biomarkers, laboratory tests and genetic data) are only exceptionally recorded and they are required more and more often in study protocols. While high quality disease registries can to some extent meet this need in specific disease areas they rarely capture co-medications, co-morbidities and adverse reactions. Improvements in the quality of in-patient care and in the recording of laboratory tests would be of value for epidemiological investigations on determinants for health outcomes, including drug-related safety issues.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Extending approved adult indications to the paediatric population is increasing and according to the European Commission's report between 50% and 90% of the medicines currently used in pediatrics have neither been tested on nor authorised for use in children (30) therefore real-world data is particularly important in this area. In our review we found that 94% of databases have some information about paediatric patients but no in-depth analysis of the available information was undertaken. A more detailed review of paediatric databases was undertaken by Neubert et al. (31) who concluded that in Europe, drug utilisation and outcome data is available for ~4 million children. However, similar to our review, the authors highlight that efforts should be made to increase availability of in-patient data, a setting where the greatest prescribing of novel medicines occurs.(31)

The majority of databases (>60%) have validity studies published. However, as highlighted by van Staa et al,(15) the systematic measurement of data quality is missing in most databases. As such and in line with the recommendations of Hall et al,(20) we encourage data holders to document the basic characteristics of their data source and to highlight when a change in recording practices occurs.
A new way forward to increase the speed and power of multi-centres studies is the use of common data models. (32) The advantage of using a CDM is that the transformed databases can be more easily integrated for research across a network. Although less than a third of databases were already transformed or in process of being transformed to a CDM in Europe, these figures are changing fast and are likely to increase due to ongoing initiatives such as the European Medical Information Network (EMIF) (33) and the future European Health Data Network (EHDN).(34)

Access to databases for research purposes can be provided at patient level in only a quarter of databases while the remaining ones had more restrictive access policies. We therefore fully support the recommendations published by other groups that governance models should be in place to facilitate data access, data sharing and secondary use of research data in health sciences. (35)

There are multiple challenges to the utilisation of EHDs in a regulatory context, particularly in Europe, which go beyond the above mentioned challenges related to the characteristics of the specific databases. These include fragmentation and lack of compatibility of resources as a result of the significant heterogeneity among European data sources, inadequate methods to integrate and analyse these heterogeneous data, the need for a systematic and consistent validation of data sources, governance issues and privacy concerns. In an attempt to deal with the significant heterogeneity

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

across data sources in Europe, ENCePP has established a Working Group dedicated to facilitating the initiation and conduct of observational research using multiple data sources.(9) As part of its work the group reviewed ongoing or finalised multi-database drug safety projects of various publicly funded EU projects which highlighted the heterogeneity of the methods used for combining EHR data from multiple databases.(3) Ongoing work of the group is centred around developing guidance on conceptual models for multi-national and multi-database studies.

Our review has a number of limitations. Firstly, we may have missed data sources during the identification process. However, we attempted to be as complete as possible by incorporating several rounds of database identification and review of the inventory by experts, including members of the ENCePP Working Group "Data sources" and database owners. Secondly, data extraction was mostly based on publicly available information and there might be inconsistencies or outdated data, even after attempts to verify the information with database owners. Thirdly, we excluded prescription only databases since they cannot be used for etiological studies; however we acknowledge their utility for drug utilisation studies which are very common in the regulatory field. Lastly, validation of the primary source data is an important process that provides confidence in the results of the analyses (36) and this was only evaluated indirectly through the number of validation studies reported by the database owners. Future work should focus on the existing validation studies and develop a more robust validation measure.

5. Conclusion

We have provided a systematic inventory of EHDs available in Europe that includes an in-depth evaluation of their capability to support regulatory decision-making on the benefits and risks of medicines in Europe. Our research showed there is a wide range of health care databases available for epidemiologic research in Europe, some of which are well-established with a long tradition in electronic recording of medical data.

The inventory has focussed on population–based electronic healthcare databases that allow causal association studies between drug exposure and health outcomes in primary care. We hope that this inventory will facilitate the identification of and access to relevant existing databases that could be used for public health research. Beyond that, we anticipate that this inventory may assist clinical

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

epidemiologists interested in undertaking other investigations such as studying the occurrence and determinants of health outcomes in a population.

We hope that this inventory will stimulate increased transparency and accessibility of other databases in addition to the development of data sources in Eastern European countries which are underrepresented.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

References

1. Good pharmacovigilance practice. Module VIII. Post-authorisation safety studies. 2016; EMA/813938/2011 Rev 3* [Accessed 2017 Nov 17]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129137.pdf

2. Santoro A, Genov G, Spooner A, Raine J, Arlett P. Promoting and Protecting Public Health: How the European Union Pharmacovigilance System Works. Drug Saf. 2017 Oct;40(10):855–69; DOI: 10.1007/s40264-017-0572-8.

3. Blake KV, Prilla S, Accadebled S, et al. European Medicines Agency review of post-authorisation studies with implications for the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. Pharmacoepidemiol Drug Saf. 2011 Oct;20(10):1021–9.

4. Mazzaglia G, Straus SMJ, Arlett P, et al. Study Design and Evaluation of Risk Minimization Measures: A Review of Studies Submitted to the European Medicines Agency for Cardiovascular, Endocrinology, and Metabolic Drugs. Drug Saf. 2017 Nov 9; DOI: 10.1007/s40264-017-0604-4.

5. Engel P, Almas MF, De Bruin ML, Starzyk K, Blackburn S, Dreyer NA. Lessons learned on the design and the conduct of Post-Authorization Safety Studies: review of 3 years of PRAC oversight. Br J Clin Pharmacol. 2017 Apr;83(4):884–93; DOI: 10.1111/bcp.13165.

6. Carroll R, Ramagopalan SV, Cid-Ruzafa J, Lambrelli D, McDonald L. An analysis of characteristics of post-authorisation studies registered on the ENCePP EU PAS Register. F1000Research. 2017 Nov 10;6:1447; DOI: 10.12688/f1000research.12198.2.

7. Goedecke T, Morales DR, Pacurariu A, Kurz X. Measuring the impact of medicines regulatory interventions - Systematic review and methodological considerations. Br J Clin Pharmacol. 2017 Dec 20; doi: 10.1111/bcp.13469.

8. Sturkenboom MCJM. Other Databases in Europe for the Analytic Evaluation of Drug Effects. Pharmacovigilance. John Wiley & Sons, Ltd; 2007. p. 361–73; DOI: 10.1002/9780470059210.ch29.

9. Kurz X, Perez-Gutthann S, ENCePP Steering Group. Strengthening standards, transparency, and collaboration to support medicine evaluation: Ten years of the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Pharmacoepidemiol Drug Saf. 2018 Jan 11; DOI: 10.1002/pds.4381

10. Klungel OH, Kurz X, de Groot MCH, et al. Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project. Pharmacoepidemiol Drug Saf. 2016 Mar;25 Suppl 1:156–65; DOI: 10.1002/pds.3968

11. Trifirò G, Coloma PM, Rijnbeek PR et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? J Intern Med. 2014 Jun 1;275(6):551–61; DOI: 10.1111/joim.12159

12. La Gamba F, Corrao G, Romio S, et al. Combining evidence from multiple electronic health care databases: performances of one-stage and two-stage meta-analysis in matched case-control studies. Pharmacoepidemiol Drug Saf. 2017 Oct;26(10):1213–9; DOI: 10.1002/pds.4280

13. Strom B. How Should One Perform Pharmacoepidemiology Studies? Choosing Among the Available Alternatives - Pharmacoepidemiology, Third Edition, Wiley Online Library, 2013, p.401-413; DOI: 10.1002/9781118344828.ch11

14. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005 Apr;58(4):323–37; DOI: 10.1016/j.jclinepi.2004.10.012

15. van Staa TP, Klungel OH. Real-life data and learning from practice to advance innovation. Available from: http://www.who.int/medicines/areas/priority_medicines/BP8_4Data.pdf

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

16. Abbing-Karahagopian V, Kurz X, de Vries F, et al. Bridging differences in outcomes of pharmacoepidemiological studies: design and first results of the PROTECT project. Curr Clin Pharmacol. 2014 May;9(2):130–8; DOI: 10.2174/1574884708666131111211802

17. Coloma PM, Schuemie MJ, Trifirò G, 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; DOI: 10.1002/pds.2053

18. Madigan D, Ryan PB, Schuemie M, et al. Evaluating the impact of database heterogeneity on observational study results. Am J Epidemiol. 2013 Aug 15;178(4):645–51; DOI: 10.1093/aje/kwt010

19. Roberto G, Leal I, Sattar N, et al. Identifying Cases of Type 2 Diabetes in Heterogeneous Data Sources: Strategy from the EMIF Project. Virgili G, editor. PLOS ONE. 2016 Aug 31;11(8):e0160648; DOI: 10.1371/journal.pone.0160648

20. Hall GC, Sauer B, Bourke A, et al. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiol Drug Saf. 2012 Jan;21(1):1–10; DOI: 10.1002/pds.2229

21. ENCePP Resources Database [Internet]. [Accessed 2018 Jan 1]. Available from: http://www.encepp.eu/encepp

22. Bridge to data [Internet]. [Accessed 2018 Jan 1]. Available from: https://www.bridgetodata.org/

23. Strom BL, editor. Pharmacoepidemiology Chichester, UK: John Wiley & Sons, Ltd; 2000.]. [Accessed 2018 Jan 1]. Available from: http://doi.wiley.com/10.1002/0470842555; DOI: 10.1002/9781118344828

24. Rothman KJ. Epidemiology: An Introduction. Oxford University Press, USA; 2002

25. PARENT Joint Action. Methodological guidelines and recommendations for efficient and rational governance of patient registries. 2015. [Accessed 2018 Jan 1]. Available from: http://patientregistries.eu/deliverables

26. Boggon R, van Staa TP, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer recording and mortality in the General Practice Research Database and linked cancer registries. Pharmacoepidemiol Drug Saf. 2013 Feb;22(2):168–75; DOI: 10.1002/pds.3374

27. Gagne JJ. Common Models, Different Approaches. Drug Saf. 2015 Aug;38(8):683–6.

28. Wettermark B, Zoëga H, Furu K, Korhonen M, Hallas J, Nørgaard M, et al. The Nordic prescription databases as a resource for pharmacoepidemiological research--a literature review. Pharmacoepidemiol Drug Saf. 2013 Jul;22(7):691–9; DOI: 10.1002/pds.3457

29. Health at a Glance: Europe 2016 | OECD READ edition [Internet]. OECD iLibrary. [Accessed 2018 Jan 15]. Available from: http://www.keepeek.com/Digital-Asset-Management/oecd/social-issues-migration-health/health-at-a-glance-europe-2016_9789264265592-en

30. EMA Annual Report 2016 [Internet]. [Accessed 2018 Jan 1] Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Annual_report/2017/05/WC500227334.pdf

31. Neubert A, Sturkenboom MCJM, Murray ML, Verhamme KMC, Nicolosi A, Giaquinto C, et al. Databases for pediatric medicine research in Europe-assessment and critical appraisal. Pharmacoepidemiol Drug Saf. 2008 Dec;17(12):1155–67; DOI: 10.1002/pds.1661

32. Gini R, Schuemie M, Brown J, et al. Data Extraction and Management in Networks of Observational Health Care Databases for Scientific Research: A Comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies. EGEMS Wash DC. 2016;4(1):1189; DOI: 10.13063/2327-9214.1189

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

33. European Medical Information Framework [Internet]. [Accessed 2018 Jan 1]. Available from: http://www.emif.eu/

34. European Health Data Network. 2018. [Internet]. [Accessed 2018 Jan 1]. Available from: http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/imi2-2017-12-04.html

35. Burton PR, Banner N, Elliot MJ, Knoppers BM, Banks J. Policies and strategies to facilitate secondary use of research data in the health sciences. Int J Epidemiol. 2017 Dec 1;46(6):1729–33; DOI: 10.1093/ije/dyx195

36. Ehrenstein V, Petersen I, Smeeth L, et al. Helping everyone do better: a call for validation studies of routinely recorded health data. Clin Epidemiol. 2016;8:49–51; DOI: 10.2147/CLEP.S104448

List of figures' and table's titles

Figure 1. Scoring of the usefulness of electronic healthcare databases available in Europe for the benefit-risk evaluation of medicines.

Figure 2. Flowchart of database selection

Table 1. List of data sources retained in the final inventory (by year)

Table 2. Distribution of data sources type and type of care covered

Figure 3. Number of data sources across Europe and duration of data collection

Footnotes

Contributions: All authors were involved in the study design and data collection. AP, XK, and DM performed the analysis and interpretation of results. All authors contributed to writing the manuscript and approved the final draft.

Funding: No sources of funding were used to assist in the preparation of this article.

Copyright statement: The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non-exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd to permit this article (if accepted) to be published in BMJ editions and any other BMJPGL products and sublicenses such use and exploit all subsidiary rights, as set out in our licence.

Conflict of interest: We have read and understood the BMJ Group policy on declaration of interests and declare the following interests: None

Disclaimer: The views expressed in this article are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the European Medicines Agency or one of its committees or working parties.

Ethical approval: Not required.

Data sharing: An extended version of the dataset is available as supplementary material.

Transparency: AP affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

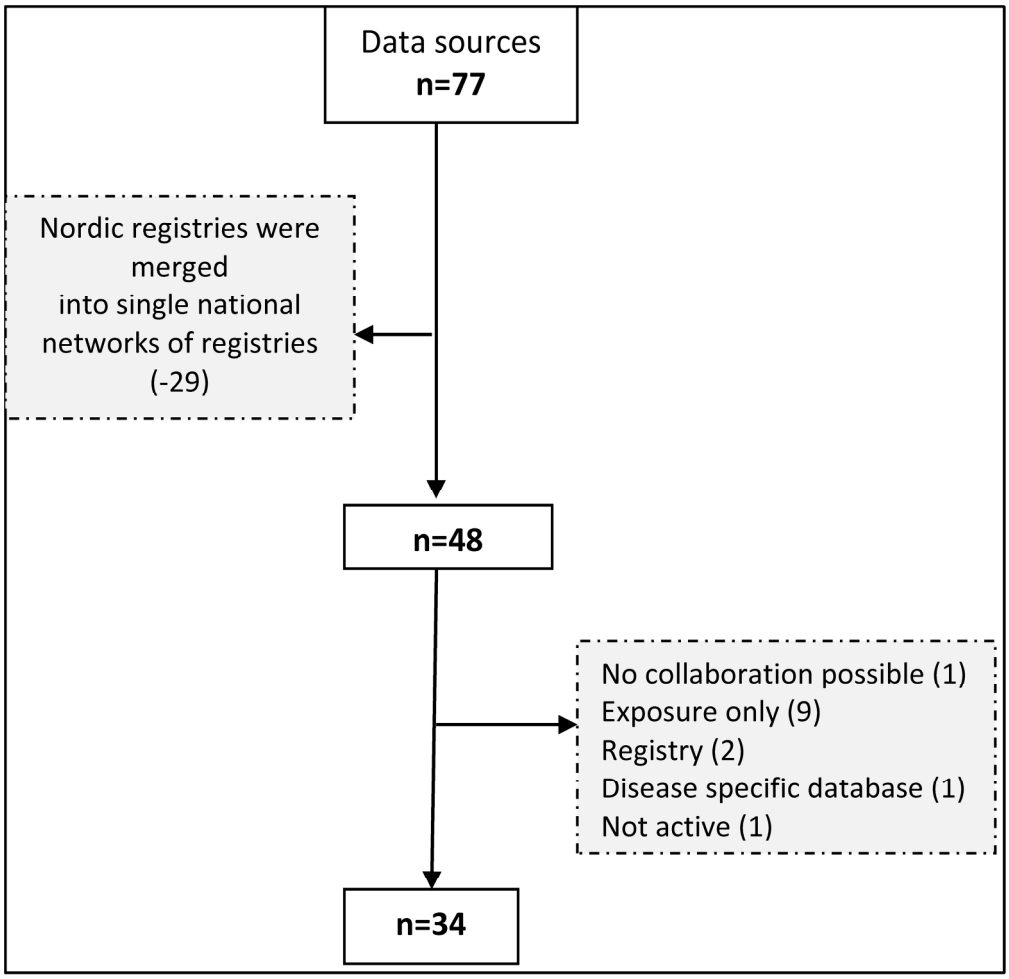
Patient involvement statement: This descriptive analysis did not involve any patients.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

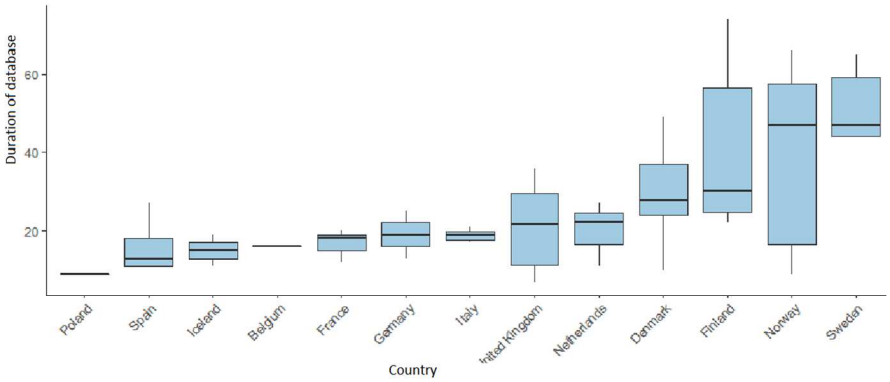
Open Access: This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

For peer review only

128x119mm (600 x 600 DPI)



102x99mm (600 x 600 DPI)



332x445mm (96 x 96 DPI)

Data source name	Abbreviation	Country	Included/Excluded	Exclusion criteria	Type	Type of care	Start date	Prescribed medicines	Dose	Duration	Route of administration	Indications	Diagnosis	Number of patients	Number of active patients	Coverage	Access and analysis of data	Linkage	Paediatric data	Hospital data	Age and/or date of birth of patients	Gender of patients	Parent-child linkage	Inclusion of vaccinations	Screening results eg scans etc	Test results (eg blood tests, blood pressure, etc)	Studies available	Is data transformed or being transformed to CDM
								0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	1 - (<5M);	1 - (<5M);		0 - No access 1 - Indirect access through atabase	0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	0 - if no assessment of	0 - if no assessment of	0 - if no assessment	0 - if no assessment of	0 - if no assessment of	0 - if no assessment of	0 - if no studies	0 - No 1 - Yes
QuintilesIMS LifeLink: Hospital Disease Database - Belgium	HDD	Belgium	Included		Electronic healthcare	Secondary care	2001	1	1	1	0	1	0	1	1	25%	3	0	1	0	1	1	0	0	1	0	0	0
UK-CRIS- Uk Case Record Interactive search	UK Cris	United Kingdom	Excluded	Disease specific database																								
IMS LifeLink:Longitudinal Prescription Data (LRx) - Belgium		Belgium	Excluded	Exposure only												30%												
Danish National and regional registries		Denmark	Included		Record linkage system	Mixed	1977	1	1	1	1	1	1	2	1	100%	1	1	1	1	1	1	0	1	1	1	2	0
Finnish National registries		Finland	Included		Record linkage system	Mixed	1964	1	1	1	1	0	1	3	3	100%	2	1	1	1	1	1	1	1	0	0	2	0
Longitudinal Prescription Data France	IMS LifeLink Treatment Dynamics	France	Excluded	Exposure only												33%												
Securite Sociale de l'Assurance Maladie	SNIIRAM	France	Included		Claims	Mixed	1999	1	1	1	0	0	1	3	3	90%	3	1	1	0	1	1	0	1	1	1	0	0
QuintilesIMS Disease Analyser	DA France	France	Included		Electronic medical records	Primary care	1997	1	1	1	0	1	1	2	1	2%	3	0	1	0	1	1	0	0	1	1	2	1
Echantillon Généraliste de Bénéficiaires	EGB	France	Included		Claims	Mixed	2006	1	1	1	0	0	1	1	1	1.03%	3	1	1	0	1	1	0	1	1	1	0	0
QuintilesIMS Disease Analyser	DA Germany	Germany	Included		Electronic medical records	Mixed	1992	1	1	1	0	1	1	3	2	3.20%	3	0	1	0	1	1	0	1	1	1	2	1
German Pharmacoepidemiological Research Database	GePaRD	Germany	Included		Claims	Mixed	2004	1	1	1	1	0	1	3	2	17%	1	1	1	1	1	1	1	1	0	0	1	0
Icelandic Registries		Iceland	Included		Record linkage system	Mixed	1998	1	1	1	1	1	1	1	1	100%	3	1	1	1	1	1	0	0	0	0	0	0
Health Service Executive - Primary Care Reimbursement Services	HSE	Ireland	Excluded	Exposure only												100%												
Pedianet Database	PediaNet	Italy	Included		Electronic medical records	Primary care	1998	1	1	1	1	1	1	1	1	3%	1	1	1	1	1	1	0	1	1	1	2	1
Agencia Regionale di Sanita Tuscany database	ARS	Italy	Included		Claims	Secondary care	1996	1	0	0	0	1	1	1	1	100%	Unknown	0	1	1	1	1	0	1	1	1	1	1
Hospital Information System -Lazio	HIS	Italy	Included		Claims	Secondary care	1998	0	0	0	0	1	0	1	1		1	1	1	1	1	1	0	0	0	0	0	0
Lombardia Health Database		Italy	Included		Electronic medical records	Secondary care	2000	1	0	0	0	1	1	2	1	100%	Unknown	1	1	1	1	1	0	0	0	0	0	0
QuintilesIMS LPD Health Search Database Longitudinal	LPD Italy	Italy	Included		Electronic medical records	Primary care	2000	1	1	1	1	1	1	1	1		3	0	0	0	1	1	0	Unknown	Unknown	1	1	0
Region Emilia Romagna Database	RER	Italy	Included		Claims	Secondary care	1997	1	Unknown	Unknown	Unknown	Unknown	1	1	1		Unknown	1	Unknown	1	1	1	0	Unknown	Unknown	Unknown	0	0
OSMED National Drug Consumption Database	OSMED	Italy	Excluded	Exposure only																								
National Health Insurance Fund Database		Lithuania	Excluded	Exposure only												40%												
IMS LifeLink - Longitudinal Prescription Data Netherlands	MS Xtrend Dynamics	Netherlands	Excluded	Exposure only												75%												
Integrated Primary Care Information Database	IPCI	Netherlands	Included		Electronic medical records	Primary care	1995	1	1	1	0	1	1	1	1		1	0	1	1	1	1	0	1	1	1	1	1
VEKTIS	VEKTIS	Netherlands	Included		Claims	Mixed	2006	1	Unknown	Unknown	Unknown	0	1	3	2	100%	2	1	1	Unknown	1	1	0	Unknown	Unknown	Unknown	Unknown	0
Pharmo Database Network	PHARMO	Netherlands	Included		Record linkage system	Mixed	1990	1	1	1	1		1	1	1	25%	1	1	1	0	1	1	0	1	0	0	2	0
Norwegian Registries		Norway	Included		Record linkage system	Mixed	1997	1	1	1	0	1	1	2	2		2	1	1	1	1	1	0	0	0	0	1	0
National Health Fund		Poland	Included		Claims	Mixed	2008	1	1	1	1	0	1	3	3	86%	1	0	1	1	1	1	0	0	0	0	0	0
Programme médicalisé des systèmes d'informations		France	Excluded	No collaboration																								
National Drug Consumption Database (Out-patient Prescribing of Drugs)		Slovenia	Excluded	Exposure only																								
Spanish National Drug Consumption Database		Spain	Excluded	Exposure only												74%												
Osservatorio Interaziendale per la Farmacoepidemiologia e la Farmacoeconomia		Italy	Excluded	Not active																								
The Information System for the Development of Research in Primary Care	SIDIAP	Spain	Included		Electronic medical records	Primary care	2006	1	1	1	0	1	1	2	2		2	1	1	1	1	1	1	1	1	1	2	1
Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria	BIFAP	Spain	Included		Electronic medical records	Primary care	2002	1	1	1	1	1	1	2	1	17%	1	0	1	0	1	1	0	0	0	1	0	0
Information System of Parc de Salut del Mar	IMASIS	Spain	Included		Electronic medical records	Primary care	1990	1	1	1	0		1	1	1		Unknown	1	1	0	1	1	0	0	0	1	0	0
JRC EUROCAT Central Registry		Europe	Excluded	Registry																								
QuintilesIMS LPD Health Search Database Longitudinal	LPD Spain	Spain	Included		Electronic medical records	Mixed	2006	1	1	1	0	1	1	1	1		3	0	1	0	1	1	0	0	0	1	0	0
Swedish National Registries		Sweden	Included		Record linkage system	Mixed	1970	1	1	1	0	0	1	1	1	98%	2	1	1	1	1	1	0	0	0	0	1	0
Clinical Practice Research Datalink - Primary care	CPRD	United Kingdom	Included		Electronic medical records	Primary care	1987	1	1	1	1	0	1	3	2		3	1	1	1	1	1	1	1	1	1	2	1
QResearch	Qresearch	United Kingdom	Included		Electronic medical records	Primary care	1989		1	1	1	1	1	3	3	43%	1	1	1	1	1	1			1	1	2	0
The electronic Data Research and Innovation Service	eDRIS-ISD	Scotland	Included		Record linkage system	Mixed	1981	1	1	1	1	0	1	2	2	100%	2	1	1	1	1	1	1	1	0	0	2	0
Secure Anonymised Information Linkage	SAIL	Wales	Included		Record linkage system	Mixed	2007	1	1	1	Unknown	1	1	3	3	100%	1	1	1	1	1	1	1	Unknown	Unknown	Unknown	0	0
EUROmediCAT	EUROmediCAT	Europe	Excluded	Registry																								
Hospital Treatment Insights	HTI	United Kingdom	Included		Record linkage system	Secondary care	2010	1	1	1	0	0	1	1	1		Unknown	1	1	1	1	1	0	0	1	1	0	0
The Health Improvement Network - Primary care	THIN	United Kingdom	Included		Electronic medical records	Primary care	2002	1	1	1	1	1	1	3	2		3	1	1	1	1	1	1	0	1	1	2	1
QuintilesIMS LPD Health Search Database Longitudinal	LPD France	France	Included		Electronic medical records	Mixed	2001	1	1	1	0	1	1	2	2		3	0	1	0	1	1	0	0	1	1	0	0
Caserta Database	Caserta	Italy	Included		Claims	Primary care	2002	1	0	1	1	1	1	1	1	100%	1	1	1	0	1	1	0	0	1	1	0	0
Medicines Monitoring Unit Scotland	MEMO	Scotland	Included		Record linkage system	Mixed	1990	1	1	1	0	Unknown	1	1	1	20%		1	1	1	1	1	Unknown	Unknown	1	1	Unknown	0
Estonian Health Insurance Fund	EHIF	Estonia	Excluded	Exposure only	Claims	Primary care	2010	1	1	1	0	1	0	1	1	95%	Unknown	1	1	0	1	1	0	1	1	1	0	0

For peer review only

BMJ Open

Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-023090.R1
Article Type:	Research
Date Submitted by the Author:	30-May-2018
Complete List of Authors:	Pacurariu, Alexandra; European Medicines Agency, Pharmacovigilance and Epidemiology Department Plueschke, Kelly; Pharmacovigilance and Epidemiology McGettigan, Patricia; Queen Mary University of London, William Harvey Research Institute; European Medicines Agency Morales, Daniel; European Medicines Agency (EMA), Surveillance & Epidemiology Service; University of Dundee, Division of Population Health Sciences Slattery, Jim; European Medicines Agency, Surveillance & Epidemiology Service Vogl, Dagmar; European Medicines Agency Goedecke, Thomas; European Medicines Agency Kurz, Xavier; European Medicines Agency, Cave, Alison; European Medicines Agency, Pharmacovigilance and Epidemiology Department
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Health policy
Keywords:	electronic healthcare databases, post-authorisation studies, regulatory science, benefit-risk evaluation, real-world data

SCHOLARONE™
Manuscripts

22nd May 2018

Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation

Alexandra Pacurariu*[1,2], Kelly Plueschke[1], Patricia McGettigan[1,3], Daniel R. Morales[1,4], Jim Slattery[1], Dagmar Vogl[1], Thomas Goedecke[1], Xavier Kurz[1], Alison Cave[1]

1 Surveillance & Epidemiology Service, European Medicines Agency (EMA), 30 Churchill Place, London E14 5EU, United Kingdom

2 Dutch Medicines Evaluation Board, 3531 AH Utrecht, the Netherlands

3 William Harvey Research Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK

4 Division of Population Health Sciences, University of Dundee, Dundee, United Kingdom

* Corresponding author: Alexandra Pacurariu alexandra.pacurariu@ema.europa.eu, +44(0)2036608559

Word count:

Number of figures and tables:

3 figures and 2 tables in the manuscript

Key words:

Electronic healthcare databases; post-authorisation studies; regulatory science; benefit-risk evaluation; real-world data; epidemiology.

For peer review only

Abstract

Objective

Electronic healthcare databases (EHDs) are useful tools for drug development and safety evaluation but their heterogeneity of structure, validity and access across Europe complicates the conduct of multi-database studies. In this paper, we provide insight into available EHDs available to support regulatory decisions on medicines.

Methods

EHDs were identified from publicly available information from the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) resources database, textbooks and web-based searches. Databases were selected using criteria related to accessibility, longitudinal dimension, recording of exposure and outcomes, and generalizability. Extracted information was verified with the database owners.

Results

A total of 34 EHDs were selected after applying key criteria relevant for regulatory purposes. The most represented regions were Northern, Central and Western Europe. The most frequent types of data source were electronic medical records (44.1%) and record linkage systems (29.4%). The median number of patients registered in the 34 data sources was 5 million (range 0.07-15 million) while the median time covered by a database was 18.5 years. Pediatric patients were included in 32 databases (94%). Completeness of information on drug exposure was variable. Published validation studies were found for only 17 databases (50%). Some level of access exists for 25 databases (73.5%) and 23 databases (67.6%) can be linked through a personal identification number to other databases with parent-child linkage possible in 7 (21%) databases. Eight databases (23.5%) were already transformed or were in the process of being transformed into a common data model.

Conclusion

The number of European database that meet minimal regulatory requirements and are readily available to be used in regulatory context is low, as well as the number of patients covered. Accessibility and validity information can be improved. This study confirmed the fragmentation, heterogeneity and lack of transparency existing in many European electronic healthcare databases.

For peer review only

DRAFT

Strengths and limitations of this study

- Data extraction was based on information provided by database owners as well as as on publicly available information
- Incomplete data capture cannot be excluded, especially for very small and new databases.
- Validation of the data source was only evaluated indirectly through the validation studies reported by the database owners
- The inventory was endorsed by an expert working group of the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (the ENCePP Working Group "Data Sources")

For peer review only
DRAFT

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1. Introduction

The European Union (EU) medicines regulatory network has responsibility for protecting patients by ensuring continuous evaluation of the safety of authorised medicines. At the core of such review is the scientific assessment of all available evidence including relevant information from the literature, results from non-clinical studies, randomised clinical trials (RCTs), observational studies, spontaneous reports and results of other available research. A way to collect more information about a medicine's safety post-marketing is by means of post-authorisation safety studies (PASS) (1). PASS may be imposed on a marketing authorisation holder by a regulatory authority or conducted by the company to address a safety concern or evaluate the effectiveness of risk minimisation measures aimed at reducing the occurrence or severity of an adverse reaction (2,3).

Secondary use of routinely collected data from electronic healthcare databases (EHDs) is often used in such studies because it is usually faster and cheaper than primary data collection.

A review of pharmaceutical industry-sponsored studies evaluating the effectiveness of risk minimisation measures submitted to the European Medicines Agency (EMA) for cardiovascular, endocrinology or metabolic drugs authorised between 1995 and 2015 found that EHDs were used in 53% of studies evaluating routine risk minimisation measures and 31% of studies evaluating additional risk minimisation measures (4). A second review of 189 PASS assessed by the EMA between 2012 and 2015 and registered in the European Union electronic Register of Post-Authorisation Studies (EU PAS Register®) showed that secondary use of routinely collected data occurred in 33.3% of them and 58% among these leveraged electronic health records (EHRs) (5). A third review of a different set of studies registered in the EU PAS Register® as of December 2016 found that 117 studies (37%) used an existing claims or electronic medical records database (6). A fourth review evaluating studies which measured the impact of regulatory interventions found that claims databases were used in 45% of studies, while EHRs were used in 22% of them, the latter being the most utilised type of data sources for such studies (7). The frequent use of EHDs in observational studies was also reported in a wider context in a review of the abstracts of presentations made at the International Conference for Pharmacoepidemiology: 53% (in 2000) and 51% (in 2005) of submitted EU pharmacoepidemiological studies were conducted using automated general practice, pharmacy or claims data (8).

The fact that between 30%-50% of observational post-authorisation studies use EHDs as their main data source reflects the importance of these data sources to support regulatory decision making (1,9). On the other hand, the use of EHDs in pre-authorisation research is currently limited and mostly focused on providing historical control data or understanding the natural history of the disease.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

As regulatory decisions based on EHDs may have a considerable impact on public health, the quality of the information, the validity and reproducibility of the derived results require close attention, especially when combining data from several data sources or when the original data is transformed before analysis (10–12). It has been emphasised that the same level of scientific rigour should be employed irrespective of the study design and data source to be used, and that the strengths and weaknesses of each data source should be considered (13). The speed at which the results could be generated is an additional important consideration, particularly for regulatory purpose (9,14,15). By considering the characteristics of the data sources and the research objectives to be addressed, the investigators should be able to choose the most appropriate resource(s) to address the question at hand. However, while some authors provide a detailed description of the databases used in their study (16–19), their description in the published literature is often incomplete and a justification for their choice in the context of alternative data sources is rarely provided (18). The International Society of Pharmacoepidemiology (ISPE) has developed guidelines to support the selection and use of data sources for observational research by highlighting potential limitations of databases and recommending testing procedures. The guidelines also provide a checklist covering six areas: database selection, use of multiple data resources, extraction and analysis of the study population, privacy and security, quality and validation procedures and documentation (20). The availability of an inventory of European databases describing the main characteristics, conditions of access and validation performed would support investigators to identify databases suitable for their research question. Moreover, knowledge of the characteristics of the data sources used in a post-authorisation study would enhance regulators' confidence in the evidence derived from such data and ultimately in the usefulness of the study in the decision-making process (21,22).

The main objective of this study is to provide an inventory of EHDs and describe their key characteristics and availability with the aim to support stakeholders in their choice of the data source when conducting a post authorisation study.

2. Methods

2.1. Identification of EHDs

As a first step, we identified existing EHDs in Europe by screening the following sources: the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) resources database (21), web-based search engines dedicated to healthcare databases (22), textbooks on clinical pharmacoepidemiology (23,24), publicly available inventories created for European Commission funded

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

research projects and databases used in publicly funded research such as EMA funded post-authorisation studies. The different population registries within the same country were considered as a single national EHD if they could be linked using a unique identification number (e.g. in Nordic countries and in Scotland).

As a second step, data sources were included in the inventory based on the following criteria: the data is available to regulatory authorities or to third-parties for research purposes; the database contains information on both drug exposure and health outcomes and is not disease or product specific; there is longitudinal data capture. Provision of relevant data for benefit-risk decision-making was one of the key criteria for selecting studies meeting regulatory requirements.

Prescription-only databases were excluded because they cannot be used for etiological studies. Product or disease specific registries were considered out of scope as they create cohorts of patients whose entry is defined either by exposure to a product or by occurrence of a disease or health outcome (25). Product specific registries are frequently used for the benefit-risk monitoring of specific products, they rarely cover a wide range of medicines and health conditions and the same registry cannot be used to monitor medicinal products of different therapeutic classes. Databases where the data collection ceased and the historical data is not accessible were excluded.

2.2. Data extraction and classification

For each database, publicly available information (on the databases' websites or in publications) was supplemented by contacting data source owners in writing. A total of 82% database owners responded. Teleconferences with seven database owners were conducted to clarify some of the information provided.

The information was extracted by six EMA reviewers (AP, KP, PMG, DM, JS, AC) and the entries for each database were cross-checked for consistency by a second reviewer. Uncertainties about the classification of any variable were resolved through discussion.

The data sources were classified in three categories according to their structure, purpose and type of data: electronic medical records, claims databases and healthcare record linkage systems (e.g. several databases are linked to form a complete database). The size of the data source was quantified by the cumulative number of patients included (both total and active patients) and number of years since the initiation of data collection in the database.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Data collected in the following categories was also recorded: demographic information (age and gender of each individual), information on prescribed or dispensed medicines (including name, dose, duration, route of administration and therapeutic indication), immunisations, diagnosis data and referrals for laboratory investigations, imaging, and other procedures. Information of laboratory tests results was not collected.

2.3. Availability of validation studies

Database owners were asked to report validation studies which they were aware of for their database. Studies published up to September 2016 were included. For the purpose of this study, a validation study was defined as any study published in a peer-reviewed journal that aimed to validate the information available on an outcome or exposure in comparison with gold standard information, usually the patients' original health records as reviewed by a medical professional or the same information captured by another database for a different purpose. For example, a study in 2012 compared records of cancer in a general practitioners' database, hospital records and cancer registries and found considerable discrepancies in cancer recording between these different data sources (26).

2.4. Accessibility

The accessibility of databases for research purposes was classified in four categories: no access, indirect access through the database owner or a third party, direct access restricted to specific datasets and direct access to the full dataset.

2.5. Coding of database characteristics for usefulness for medicines benefit-risk evaluation

Instead of evaluating the quality of each database, we aimed to assist in the selection of databases by implementing a coding process that identifies the data sources considered to provide sufficient information to contribute to regulatory questions on the benefit-risk evaluation of medicines. For this purpose, the following domains were included in the coding process: extent of data capture of study variables, size of data source, quality and validity of information, accessibility, potential for linkage and existing process in place to convert the data into a common data model (CDM) (Figure 1). A common data model provides a common representation and architecture of the data across multiple databases, thus enabling the standardisation of administrative and clinical information and allowing the use of common analytical tools (27).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

[Insert Figure 1. Coding of the characteristics of electronic healthcare databases available in Europe for the benefit-risk evaluation of medicines.]

The ENCePP Working Group "Data Sources" (9) reviewed an initial version of the inventory with the description of databases and endorsed the final inventory.

Patient involvement statement: This descriptive analysis did not involve any patients.

3. Results

3.1. General overview

The initial search generated a list of 77 potential data sources. After merging the national registries into a single entry and applying the exclusion criteria, 34 of them were retained in the final inventory (Figure 2). Table 1 provides a list of these 34 databases and the complete information is provided in the *Supplementary material.*

[Insert Figure 2. Flowchart of database selection]

DRAFT

For peer review only

Table 1. List of data sources retained in the final inventory (by year)

Data source name	Country	Type	Type of care	Start date
Finnish National registries	Finland	Record linkage system	Mixed	1964
Swedish National registries	Sweden	Record linkage system	Mixed	1970
Danish National and regional registries	Denmark	Record linkage system	Mixed	1977
The electronic Data Research and Innovation Service	Scotland	Record linkage system	Mixed	1981
Clinical Practice Research Datalink	United Kingdom	Electronic medical records	Primary care	1987
QResearch	United Kingdom	Electronic medical records	Primary care	1989
Information System of Parc de Salut del Mar	Spain	Electronic medical records	Primary care	1990
Medicines Monitoring Unit Scotland	Scotland	Record linkage system	Mixed	1990
Pharmo Database Network	Netherlands	Record linkage system	Mixed	1990
QuintilesIMS Disease Analyser	Germany	Electronic medical records	Mixed	1992
Integrated Primary Care Information Database	Netherlands	Electronic medical records	Primary care	1995
Agencia Regionale di Sanita Tuscany database	Italy	Claims	Secondary care	1996
Norwegian Registries	Norway	Record linkage system	Mixed	1997
QuintilesIMS Disease Analyser	France	Electronic medical records	Primary care	1997
Region Emilia Romagna Database	Italy	Claims	Secondary care	1997
Hospital Information System -Lazio	Italy	Claims	Secondary care	1998
Icelandic Registries	Iceland	Record linkage system	Mixed	1998
Pedianet Database	Italy	Electronic medical records	Primary care	1998
Securite Sociale de l'Assurance Maladie	France	Claims	Mixed	1999
Lombardia Health Database	Italy	Electronic medical records	Secondary care	2000
QuintilesIMS LPD Health Search Database Longitudinal	Italy	Electronic medical records	Primary care	2000
QuintilesIMS LifeLink: Hospital Disease Database - Belgium	Belgium	Electronic medical records	Secondary care	2001
QuintilesIMS LPD Health Search Database Longitudinal	France	Electronic medical records	Mixed	2001
Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria	Spain	Electronic medical records	Primary care	2002
Caserta Database	Italy	Claims	Primary care	2002
The Health Improvement Network	United Kingdom	Electronic medical records	Primary care	2002
German Pharmacoepidemiological Research Database	Germany	Claims	Mixed	2004
Echantillon Généraliste de Bénéficiaires	France	Claims	Mixed	2006
QuintilesIMS LPD Health Search Database Longitudinal	Spain	Electronic medical records	Mixed	2006
The Information System for the Development of Research in Primary Care	Spain	Electronic medical records	Primary care	2006
VEKTIS	Netherlan	Claims	Mixed	2006

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	ds			
Secure Anonymised Information Linkage	Wales	Record linkage system	Mixed	2007
National Health Fund	Poland	Claims	Mixed	2008
Hospital Treatment Insights	United Kingdom	Record linkage system	Secondary care	2010

The most frequent types of data source identified were electronic medical records (n=15, 44.1%) followed by record linkage systems (n=10, 29.4%) and claims databases (n=9, 26.5%). In terms of the type of care covered, mixed care settings (primary and secondary care) were most common (n=17, 50%), followed by primary care databases (n=11, 32.3%) (Table 2). The median number of patients followed cumulatively across the 34 data sources was 5 million (range 0.07-15 million).

Table 2. Distribution of data sources type and type of care covered

Type of data source	Primary care	Secondary care	Mixed
Claims	1 (2.9%)	3 (8.82%)	5 (14.7%)
Electronic medical records	10 (29.4%)	2 (5.9%)	3 (8.8%)
Record linkage system	0 (0%)	1 (2.9%)	9 (26.5%)

Patient age and gender were recorded in all data sources while pediatric patients were included in 32 databases (94%). The median year for database start was 1998, with the oldest database established in 1964 (the Finnish Hospital Discharge Register). The median calendar time covered by a database was 18.5 years (range 7-53 years). In terms of geographical coverage, 17% of databases collect data from Norway, 14% from Finland and 10% from Denmark and Italy (Figure 3).

[Insert
Figure 3: European data sources and duration of data collection.

3.2. Information captured

By definition, all the databases retained in the final inventory contained information about drug exposure (either prescribed or dispensed). The completeness of information was however variable: 28 databases (82.3%) had information about prescribed dose and duration of treatment (either directly recorded or inferred from other collected variables); 14 (41.1%) had information about route of administration; 20 databases (58.8%) recorded the therapeutic indication associated with the prescription (either directly recorded or inferred from other database elements). Over-the-counter

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

drugs were not captured in any of the databases while vaccinations were captured in 13 databases (38.0%). Data on hospital in-patient administered drugs were rarely captured (5.8%).

All databases had information about medical events (diagnosis) as a prerequisite for inclusion in our inventory. Referrals for laboratory investigations were captured in 19 (55.9%) and referrals for imaging or other diagnostic procedures were captured in 16 databases (47.1%).

3.3. Validation studies

No published validation study was reported for 17 databases (50.0%), while a total of 42 validation studies were reported for the other 17 databases, with a median of 3 validation studies per database (range: 1-25). The validation concerned either specific health outcomes or prescription information. The most common gold standards used for the validation included paper based prescriptions, medical records, death records and perinatal deaths obtained from registries or national statistics reports. Some database owners have reported as validation studies the validation of prediction algorithms for various health outcomes as chronic kidney disease, ischaemic stroke and various types of cancers based on an estimating the absolute risk of a particular outcome in primary care patients with and without symptoms (1,2). It is debatable if these are truly validation studies according to our definition.

3.4. Accessibility and potential for linkage

One database was excluded due to lack of access to third parties, 10 (29.4%) offer indirect access to the database for third parties, 7 (20.5%) provide direct access to specific datasets and 8 (23.5%) offer direct access to the full dataset. The level of access could not be identified for 8 EHDs (23.5%). In terms of linkage, 23 databases (67.6%) could be linked through a unique personal identification number (PIN) to other databases containing additional healthcare-related information including cause of death registries, hospital data, prescription databases and cancer registries. The Nordic registries are a good example of extensive linkage among different national registries through usage of a PIN. Other forms of linkage exist. In order to avoid the use of PIN and preserve anonymity, the PHARMO network uses probabilistic linkage based on patient birth date, gender and general practitioner code. The linkage of a parent with their child ('parent-child linkage'), which is useful for studies investigating pregnancy exposures and effect on offspring, was available in 7 data sources (20.6%).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

3.5. Conversion of the database to a CDM

Four (11.7%) databases were already transformed in a CDM and four others were in the process of being converted to a CDM (the QuintilesIMS Disease Analyser France and Germany, the Spanish Information System for the Development of Research in Primary Care, the Agenzia Regionale di Sanità Tuscany database, The Pedianet, the Clinical Practice Research Datalink, the Integrated Primary Care Information Database and The Health Improvement Network). Seven of these 8 databases used the Observational Medical Outcomes Partnership CDM (27), while the Spanish Information System for the Development of Research in Primary Care is implementing the model used in the ADVANCE project for vaccine studies (28).

4. Discussion

A total of 34 European EHDs with potential for use in the regulatory environment were included in this study. The regions most frequently represented were Northern, Central and Western Europe, with a scarcity of data sources in Eastern Europe. The most common data sources assessed were electronic medical records with a mix of primary and secondary care coverage. Most of the databases contain out-patient prescribing while in-patient prescribing is very rarely captured. The median number of patients registered within the 34 data sources was 5 million and the median calendar time covered by a database was 18.5 years. In terms of accessibility, 24% of databases offered direct access to the full data source, with the rest having a somewhat more limited access. There are a few similar studies of EHDs available in Europe (8,29), but as far as we are aware this is the first study taking a regulatory perspective. An analysis of the characteristics of post-authorisation studies requested by regulators showed that 47% of studies involved secondary use of data emphasizing the important role of secondary data in the regulatory setting. More detailed descriptions of database characteristics are provided in electronic repositories such as the European Medical Information Network (EMIF), the ENCePP resource database and the Bridge to Data initiative (21,22). However existing repositories are either incomplete, have a limited coverage or they require a fee for access, therefore restricting access to their information.

This study helps identify databases with key characteristics as an entry door to further investigate with their owner their potential usefulness for a specific study.

Given that different national guidelines and clinical practice can generate significant heterogeneity in how healthcare is delivered and recorded (30), it is important that regulators have access to data from

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

as broad a geographical spread as possible. Thus there is a clear need for the development of data sources in EU member states which currently either have no data sources or are poorly represented. The data recorded in the databases include some limitations. Firstly, the limited capture of in-patient prescribing poses a problem for regulators and investigators since many newly approved drugs are specialised drugs, used exclusively in secondary care (31). Secondly, some disease-specific variables (e.g., biomarkers, laboratory tests and genetic data) are only exceptionally recorded and they are required more and more often in study protocols. High quality disease registries can to some extent meet this need in specific disease areas but they rarely capture co-medications, co-morbidities and adverse reactions. Improvements in the quality of in-patient care and in the recording of laboratory tests would be of value for epidemiological investigations on determinants for health outcomes, including drug-related safety issues.

With regards to validation, 50% of databases had at least one validation study published. Validation should normally be done for the data elements collected in every study. Publication of validation studies is not an indicator of the overall validity of the database but may inform researchers on the feasibility to perform study-specific validation in a database. A repository of validated outcomes in specific databases would reduce duplication of work. Such a repository should include a clear description of the methodology and limitations of the analysis.

Extending approved adult indications to the pediatric population is increasing and according to the European Commission's report between 50% and 90% of the medicines currently used in pediatrics have neither been tested on nor authorised for use in children (31). Availability of real-world data is therefore particularly important for this purpose. In our review we found that 94% of databases have some information about pediatric patients but no in-depth analysis of the available information was undertaken. A more detailed review of pediatric databases was undertaken by Neubert et al. who concluded that in Europe, drug utilisation and outcome data is available for ~4 million children (32). However, similar to our study, the authors highlight that efforts should be made to increase availability of in-patient data, a setting where the greatest prescribing of novel medicines occurs (32).

While validity studies were published for half of the databases, van Staa *et al*. (15) highlighted that systematic measurement of data quality is lacking in most databases. As such and in line with the recommendations of Hall et al (20), we encourage data holders to document the basic characteristics of their data source and to highlight when a change in recording practices occurs.

A new way forward to increase the speed and power of multi-centres studies is the use of a CDM (33). The advantage of using a CDM is that the transformed databases can be more easily integrated for

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

research across a network. Although less than a third of databases were already converted or in process of being converted to a CDM in Europe, these figures are likely to change fast due to ongoing initiatives such as EMIF (34) and the European Health Data Network project (EHDN) (35).

Access to databases for research purposes can be provided at patient level in only a quarter of databases while the remaining ones had more restrictive access policies. We therefore fully support the recommendations published by other groups that governance models should be in place to facilitate data access, data sharing and secondary use of research data in health sciences (36).

There are multiple challenges to the utilisation of EHDs in a regulatory context, particularly in Europe, which go beyond the above mentioned challenges related to the characteristics of the specific databases. These include fragmentation and lack of interoperability of European data sources, inconsistent use of methods to integrate and analyse heterogeneous data, lack of systematic and consistent validation of data sources, governance issues and privacy concerns. In an attempt to deal with the significant heterogeneity across data sources in Europe, ENCePP has established a Working Group dedicated to facilitating the initiation and conduct of observational research using multiple data sources (9). As part of its work the group reviewed ongoing or finalised multi-database drug safety projects of various publicly funded EU projects, which highlighted the heterogeneity of the methods used for combining EHR data from multiple databases (3). Ongoing work of the group is centred around developing guidance on conceptual models for multi-national and multi-database studies.

Our review has a number of limitations. Firstly, we may have missed data sources during the identification process. However, we attempted to be as complete as possible by incorporating several rounds of database identification and review of the inventory by experts, including members of the ENCePP Working Group "Data sources" and database owners. The difficulties we encountered when trying to map all the existing EHDs in Europe highlight again the need for more comprehensive and accessible repositories with EHDs.

Secondly, we excluded prescription only databases since they cannot be used for etiological studies even if we acknowledge their utility for drug utilisation studies which are very common in the regulatory field. Lastly, validation of the primary source data is an important process that provides confidence in the results of the analyses (37) and this was only evaluated indirectly through the number of validation studies reported by the database owners. A strength of our study was that data from publicly available sources was complemented or verified with database owners.

There is more work to be done in order to increase transparency and accessibility of existing datasources. Examples of areas for future development are to develop more robust validation

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

measures and increase transparency of validated outcomes, to transform databases through a CDM to allow faster feasibility assessment and execution of studies, and to stimulate creation and access to EHRs in Eastern Europe.

5. Conclusion

We have provided a systematic inventory of EHDs available in Europe that includes a summary evaluation of their capability to support regulatory decision-making on the benefits and risks of medicines in Europe. Our research showed that despite the wide range of health care databases available for epidemiologic research in Europe, many of them were excluded due to the absence of information needed for key regulatory activities. This analysis confirmed the fragmentation, heterogeneity and lack of transparency existing in European electronic healthcare databases.

The analysis has focussed on population–based electronic healthcare databases allowing conducting causal association studies between drug exposure and health outcomes in primary care. Our intention is to help the identification of and access to relevant existing databases that could be used for public health research. Beyond this objective, we hope that this inventory may assist clinical epidemiologists interested in undertaking other investigations such as studying the occurrence and determinants of health outcomes in a population.

We hope that this inventory will stimulate increased transparency and accessibility of other databases in addition to the development of data sources in Eastern European countries, which are currently under-represented.

DRAFT
For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

References

1. Good pharmacovigilance practice. Module VIII. Post-authorisation safety studies. 2016; EMA/813938/2011 Rev 3* [Accessed 2017 Nov 17]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129137.pdf

2. Santoro A, Genov G, Spooner A, Raine J, Arlett P. Promoting and Protecting Public Health: How the European Union Pharmacovigilance System Works. Drug Saf. 2017 Oct;40(10):855–69; DOI: 10.1007/s40264-017-0572-8.

3. Blake KV, Prilla S, Accadebled S, et al. European Medicines Agency review of post-authorisation studies with implications for the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. Pharmacoepidemiol Drug Saf. 2011 Oct;20(10):1021–9.

4. Mazzaglia G, Straus SMJ, Arlett P, et al. Study Design and Evaluation of Risk Minimization Measures: A Review of Studies Submitted to the European Medicines Agency for Cardiovascular, Endocrinology, and Metabolic Drugs. Drug Saf. 2017 Nov 9; DOI: 10.1007/s40264-017-0604-4.

5. Engel P, Almas MF, De Bruin ML, Starzyk K, Blackburn S, Dreyer NA. Lessons learned on the design and the conduct of Post-Authorization Safety Studies: review of 3 years of PRAC oversight. Br J Clin Pharmacol. 2017 Apr;83(4):884–93; DOI: 10.1111/bcp.13165.

6. Carroll R, Ramagopalan SV, Cid-Ruzafa J, Lambrelli D, McDonald L. An analysis of characteristics of post-authorisation studies registered on the ENCePP EU PAS Register. F1000Research. 2017 Nov 10;6:1447; DOI: 10.12688/f1000research.12198.2.

7. Goedecke T, Morales DR, Pacurariu A, Kurz X. Measuring the impact of medicines regulatory interventions - Systematic review and methodological considerations. Br J Clin Pharmacol. 2017 Dec 20; doi: 10.1111/bcp.13469.

8. Sturkenboom MCJM. Other Databases in Europe for the Analytic Evaluation of Drug Effects. Pharmacovigilance. John Wiley & Sons, Ltd; 2007. p. 361–73; DOI: 10.1002/9780470059210.ch29.

9. Kurz X, Perez-Gutthann S, ENCePP Steering Group. Strengthening standards, transparency, and collaboration to support medicine evaluation: Ten years of the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Pharmacoepidemiol Drug Saf. 2018 Jan 11; DOI: 10.1002/pds.4381

10. Klungel OH, Kurz X, de Groot MCH, et al. Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project. Pharmacoepidemiol Drug Saf. 2016 Mar;25 Suppl 1:156–65; DOI: 10.1002/pds.3968

11. Trifirò G, Coloma PM, Rijnbeek PR et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? J Intern Med. 2014 Jun 1;275(6):551–61; DOI: 10.1111/joim.12159

12. La Gamba F, Corrao G, Romio S, et al. Combining evidence from multiple electronic health care databases: performances of one-stage and two-stage meta-analysis in matched case-control studies. Pharmacoepidemiol Drug Saf. 2017 Oct;26(10):1213–9; DOI: 10.1002/pds.4280

13. Strom B. How Should One Perform Pharmacoepidemiology Studies? Choosing Among the Available Alternatives - Pharmacoepidemiology, Third Edition, Wiley Online Library, 2013, p.401-413; DOI: 10.1002/9781118344828.ch11

14. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005 Apr;58(4):323–37; DOI: 10.1016/j.jclinepi.2004.10.012

15. van Staa TP, Klungel OH. Real-life data and learning from practice to advance innovation. Available from: http://www.who.int/medicines/areas/priority_medicines/BP8_4Data.pdf

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

16. Abbing-Karahagopian V, Kurz X, de Vries F, et al. Bridging differences in outcomes of pharmacoepidemiological studies: design and first results of the PROTECT project. Curr Clin Pharmacol. 2014 May;9(2):130–8; DOI: 10.2174/1574884708666131111211802

17. Coloma PM, Schuemie MJ, Trifirò G, 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; DOI: 10.1002/pds.2053

18. Madigan D, Ryan PB, Schuemie M, et al. Evaluating the impact of database heterogeneity on observational study results. Am J Epidemiol. 2013 Aug 15;178(4):645–51; DOI: 10.1093/aje/kwt010

19. Roberto G, Leal I, Sattar N, et al. Identifying Cases of Type 2 Diabetes in Heterogeneous Data Sources: Strategy from the EMIF Project. Virgili G, editor. PLOS ONE. 2016 Aug 31;11(8):e0160648; DOI: 10.1371/journal.pone.0160648

20. Hall GC, Sauer B, Bourke A, et al. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiol Drug Saf. 2012 Jan;21(1):1–10; DOI: 10.1002/pds.2229

21. ENCePP Resources Database [Internet]. [Accessed 2018 Jan 1]. Available from: http://www.encepp.eu/encepp

22. Bridge to data [Internet]. [Accessed 2018 Jan 1]. Available from: https://www.bridgetodata.org/

23. Strom BL, editor. Pharmacoepidemiology Chichester, UK: John Wiley & Sons, Ltd; 2000.]. [Accessed 2018 Jan 1]. Available from: http://doi.wiley.com/10.1002/0470842555; DOI: 10.1002/9781118344828

24. Rothman KJ. Epidemiology: An Introduction. Oxford University Press, USA; 2002

25. PARENT Joint Action. Methodological guidelines and recommendations for efficient and rational governance of patient registries. 2015. [Accessed 2018 Jan 1]. Available from: http://patientregistries.eu/deliverables

26. Boggon R, van Staa TP, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer recording and mortality in the General Practice Research Database and linked cancer registries. Pharmacoepidemiol Drug Saf. 2013 Feb;22(2):168–75; DOI: 10.1002/pds.3374

27. Gagne JJ. Common Models, Different Approaches. Drug Saf. 2015 Aug;38(8):683–6.

28. ADVANCE website [Accessed 2018 May 23]. Available from : http://www.advance-vaccines.eu/

29. Wettermark B, Zoëga H, Furu K, Korhonen M, Hallas J, Nørgaard M, et al. The Nordic prescription databases as a resource for pharmacoepidemiological research--a literature review. Pharmacoepidemiol Drug Saf. 2013 Jul;22(7):691–9; DOI: 10.1002/pds.3457

30. Health at a Glance: Europe 2016 | OECD READ edition [Internet]. OECD iLibrary. [Accessed 2018 Jan 15]. Available from: http://www.keepeek.com/Digital-Asset-Management/oecd/social-issues-migration-health/health-at-a-glance-europe-2016_9789264265592-en

31. EMA Annual Report 2016 [Internet]. [Accessed 2018 Jan 1] Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Annual_report/2017/05/WC500227334.pdf

32. Neubert A, Sturkenboom MCJM, Murray ML, Verhamme KMC, Nicolosi A, Giaquinto C, et al. Databases for pediatric medicine research in Europe-assessment and critical appraisal. Pharmacoepidemiol Drug Saf. 2008 Dec;17(12):1155–67; DOI: 10.1002/pds.1661

33. Gini R, Schuemie M, Brown J, et al. Data Extraction and Management in Networks of Observational Health Care Databases for Scientific Research: A Comparison of EU-ADR, OMOP,

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Mini-Sentinel and MATRICE Strategies. EGEMS Wash DC. 2016;4(1):1189; DOI: 10.13063/2327-9214.1189

34. European Medical Information Framework [Internet]. [Accessed 2018 Jan 1]. Available from: http://www.emif.eu/

35. European Health Data Network. 2018. [Internet]. [Accessed 2018 Jan 1]. Available from: http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/imi2-2017-12-04.html

36. Burton PR, Banner N, Elliot MJ, Knoppers BM, Banks J. Policies and strategies to facilitate secondary use of research data in the health sciences. Int J Epidemiol. 2017 Dec 1;46(6):1729–33; DOI: 10.1093/ije/dyx195

37. Ehrenstein V, Petersen I, Smeeth L, et al. Helping everyone do better: a call for validation studies of routinely recorded health data. Clin Epidemiol. 2016;8:49–51; DOI: 10.2147/CLEP.S104448

List of figures' and table's titles

Figure 1. Coding of the characteristics of electronic healthcare databases available in Europe for the benefit-risk evaluation of medicines.

Figure 2. Flowchart of database selection

Figure 3. European data sources and duration of data collection

Table 1. List of data sources retained in the final inventory (by year)

Table 2. Distribution of data sources type and type of care covered

Footnotes

Contributions: All authors (AP, KP,PMG, DM, JS, DV, TG, XK and AC) were involved in the study design and data collection. AP, XK, and DM performed the analysis and interpretation of results. AP, DM and AC contributed to writing and KP,PMG, JS, TG, XK revised and approved the final draft.

Funding: No sources of funding were used to assist in the preparation of this article.

Copyright statement: The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non-exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd to permit this article (if accepted) to be published in BMJ editions and any other BMJPGL products and sublicenses such use and exploit all subsidiary rights, as set out in our licence.

Conflict of interest: We have read and understood the BMJ Group policy on declaration of interests and declare the following interests: None

Disclaimer: The views expressed in this article are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the European Medicines Agency or one of its committees or working parties.

Ethical approval: Not required.

Data sharing: An extended version of the dataset is available as supplementary material.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Transparency: AP affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Open Access: This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

DRAFT
For peer review only

Figure 1. Coding of European electronic healthcare databases characteristics for the benefit-risk evaluation of medicines.
The coding system was binary: 0 if information was absent and 1 if it was present. The degree of completion for a specific variable was not recorded. An exception to the binary classification was done for the accessibility variable: 0 - no access; 1 - indirect access through database owner or third party; 2 - direct access to specific data sources; 3 - direct access to full data source.

150x150mm (300 x 300 DPI)

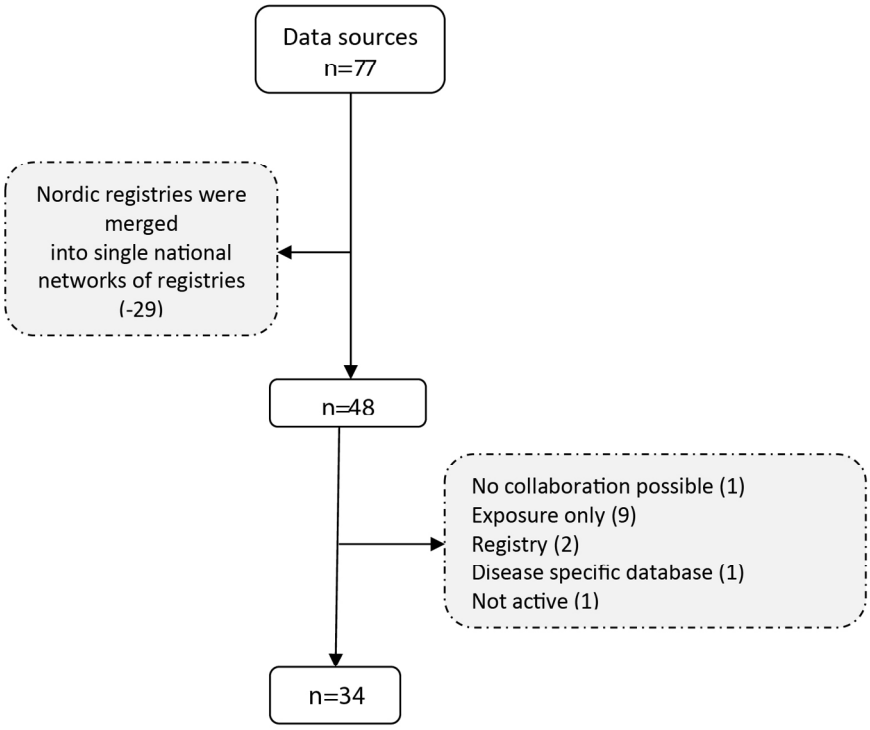


Figure 2. Flowchart of database selection

150x150mm (300 x 300 DPI)

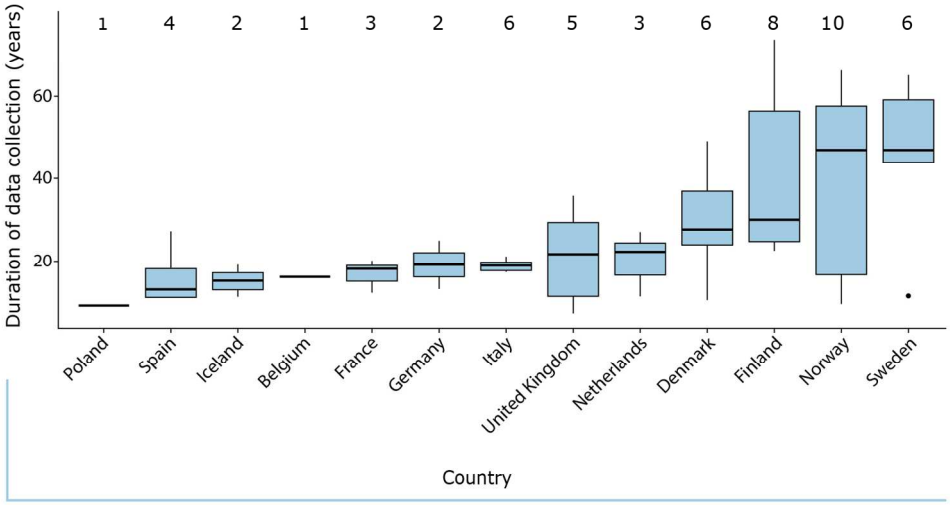


Figure 3. European data sources and duration of data collection
Boxplots indicate the median (horizontal black line) data collection time by country while the margins of the boxplot represent the interquartile range, the vertical lines indicate the minimum and maximum values. The number of databases per country are provided above the boxplots.

150x150mm (300 x 300 DPI)

Data source name	Abbreviation	Country	Included/Excluded	Exclusion criteria	Type	Type of care	Start date	Prescribed medicines	Dose	Duration	Route of administration	Indications	Diagnosis	Number of patients	Number of active patients	Coverage	Access and analysis of data	Linkage	Paediatric data	Hospital data	Age and/or date of birth of patients	Gender of patients	Parent-child linkage	Inclusion of vaccinations	Screening results eg scans etc	Test results (eg blood tests, blood pressure, etc)	Studies available	Is data transformed or being transformed to CDM
								0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	1 - (<5M);	1 - (<5M);		0 - No access 1 - Indirect access through atabase	0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	0 - if no assessment of	0 - if no assessment of	0 - if no assessment	0 - if no assessment of	0 - if no assessment of	0 - if no assessment of	0 - if no studies	0 - No 1 - Yes
QuintilesIMS LifeLink: Hospital Disease Database - Belgium	HDD	Belgium	Included		Electronic healthcare	Secondary care	2001	1	1	1	0	1	0	1	1	25%	3	0	1	0	1	1	0	0	1	0	0	0
UK-CRIS- Uk Case Record Interactive search	UK Cris	United Kingdom	Excluded	Disease specific database																								
IMS LifeLink:Longitudinal Prescription Data (LRx) - Belgium		Belgium	Excluded	Exposure only												30%												
Danish National and regional registries		Denmark	Included		Record linkage system	Mixed	1977	1	1	1	1	1	1	2	1	100%	1	1	1	1	1	1	0	1	1	1	2	0
Finnish National registries		Finland	Included		Record linkage system	Mixed	1964	1	1	1	1	0	1	3	3	100%	2	1	1	1	1	1	1	1	0	0	2	0
Longitudinal Prescription Data France	IMS LifeLink Treatment Dynamics	France	Excluded	Exposure only												33%												
Securite Sociale de l'Assurance Maladie	SNIIRAM	France	Included		Claims	Mixed	1999	1	1	1	0	0	1	3	3	90%	3	1	1	0	1	1	0	1	1	1	0	0
QuintilesIMS Disease Analyser	DA France	France	Included		Electronic medical records	Primary care	1997	1	1	1	0	1	1	2	1	2%	3	0	1	0	1	1	0	0	1	1	2	1
Echantillon Généraliste de Bénéficiaires	EGB	France	Included		Claims	Mixed	2006	1	1	1	0	0	1	1	1	1.03%	3	1	1	0	1	1	0	1	1	1	0	0
QuintilesIMS Disease Analyser	DA Germany	Germany	Included		Electronic medical records	Mixed	1992	1	1	1	0	1	1	3	2	3.20%	3	0	1	0	1	1	0	1	1	1	2	1
German Pharmacoepidemiological Research Database	GePaRD	Germany	Included		Claims	Mixed	2004	1	1	1	1	0	1	3	2	17%	1	1	1	1	1	1	1	1	0	0	1	0
Icelandic Registries		Iceland	Included		Record linkage system	Mixed	1998	1	1	1	1	1	1	1	1	100%	3	1	1	1	1	1	0	0	0	0	0	0
Health Service Executive - Primary Care Reimbursement Services	HSE	Ireland	Excluded	Exposure only												100%												
Pedianet Database	PediaNet	Italy	Included		Electronic medical records	Primary care	1998	1	1	1	1	1	1	1	1	3%	1	1	1	1	1	1	0	1	1	1	2	1
Agencia Regionale di Sanita Tuscany database	ARS	Italy	Included		Claims	Secondary care	1996	1	0	0	0	1	1	1	1	100%	Unknown	0	1	1	1	1	0	1	1	1	1	1
Hospital Information System -Lazio	HIS	Italy	Included		Claims	Secondary care	1998	0	0	0	0	1	0	1	1		1	1	1	1	1	1	0	0	0	0	0	0
Lombardia Health Database		Italy	Included		Electronic medical records	Secondary care	2000	1	0	0	0	1	1	2	1	100%	Unknown	1	1	1	1	1	0	0	0	0	0	0
QuintilesIMS LPD Health Search Database Longitudinal	LPD Italy	Italy	Included		Electronic medical records	Primary care	2000	1	1	1	1	1	1	1	1		3	0	0	0	1	1	0	Unknown	Unknown	1	1	0
Region Emilia Romagna Database	RER	Italy	Included		Claims	Secondary care	1997	1	Unknown	Unknown	Unknown	Unknown	1	1	1		Unknown	1	Unknown	1	1	1	0	Unknown	Unknown	Unknown	0	0
OSMED National Drug Consumption Database	OSMED	Italy	Excluded	Exposure only																								
National Health Insurance Fund Database		Lithuania	Excluded	Exposure only												40%												
IMS LifeLink - Longitudinal Prescription Data Netherlands	MS Xtrend Dynamics	Netherlands	Excluded	Exposure only												75%												
Integrated Primary Care Information Database	IPCI	Netherlands	Included		Electronic medical records	Primary care	1995	1	1	1	0	1	1	1	1		1	0	1	1	1	1	0	1	1	1	1	1
VEKTIS	VEKTIS	Netherlands	Included		Claims	Mixed	2006	1	Unknown	Unknown	Unknown	0	1	3	2	100%	2	1	1	Unknown	1	1	0	Unknown	Unknown	Unknown	Unknown	0
Pharmo Database Network	PHARMO	Netherlands	Included		Record linkage system	Mixed	1990	1	1	1	1		1	1	1	25%	1	1	1	0	1	1	0	1	0	0	2	0
Norwegian Registries		Norway	Included		Record linkage system	Mixed	1997	1	1	1	0	1	1	2	2		2	1	1	1	1	1	0	0	0	0	1	0
National Health Fund		Poland	Included		Claims	Mixed	2008	1	1	1	1	0	1	3	3	86%	1	0	1	1	1	1	0	0	0	0	0	0
Programme médicalisé des systèmes d'informations		France	Excluded	No collaboration																								
National Drug Consumption Database (Out-patient Prescribing of Drugs)		Slovenia	Excluded	Exposure only																								
Spanish National Drug Consumption Database		Spain	Excluded	Exposure only												74%												
Osservatorio Interaziendale per la Farmacoepidemiologia e la Farmacoeconomia		Italy	Excluded	Not active																								
The Information System for the Development of Research in Primary Care	SIDIAP	Spain	Included		Electronic medical records	Primary care	2006	1	1	1	0	1	1	2	2		2	1	1	1	1	1	1	1	1	1	2	1
Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria	BIFAP	Spain	Included		Electronic medical records	Primary care	2002	1	1	1	1	1	1	2	1	17%	1	0	1	0	1	1	0	0	0	1	0	0
Information System of Parc de Salut del Mar	IMASIS	Spain	Included		Electronic medical records	Primary care	1990	1	1	1	0		1	1	1		Unknown	1	1	0	1	1	0	0	0	1	0	0
JRC EUROCAT Central Registry		Europe	Excluded	Registry																								
QuintilesIMS LPD Health Search Database Longitudinal	LPD Spain	Spain	Included		Electronic medical records	Mixed	2006	1	1	1	0	1	1	1	1		3	0	1	0	1	1	0	0	0	1	0	0
Swedish National Registries		Sweden	Included		Record linkage system	Mixed	1970	1	1	1	0	0	1	1	1	98%	2	1	1	1	1	1	0	0	0	0	1	0
Clinical Practice Research Datalink - Primary care	CPRD	United Kingdom	Included		Electronic medical records	Primary care	1987	1	1	1	1	0	1	3	2		3	1	1	1	1	1	1	1	1	1	2	1
QResearch	Qresearch	United Kingdom	Included		Electronic medical records	Primary care	1989		1	1	1	1	1	3	3	43%	1	1	1	1	1	1			1	1	2	0
The electronic Data Research and Innovation Service	eDRIS-ISD	Scotland	Included		Record linkage system	Mixed	1981	1	1	1	1	0	1	2	2	100%	2	1	1	1	1	1	1	1	0	0	2	0
Secure Anonymised Information Linkage	SAIL	Wales	Included		Record linkage system	Mixed	2007	1	1	1	Unknown	1	1	3	3	100%	1	1	1	1	1	1	1	Unknown	Unknown	Unknown	0	0
EUROmediCAT	EUROmediCAT	Europe	Excluded	Registry																								
Hospital Treatment Insights	HTI	United Kingdom	Included		Record linkage system	Secondary care	2010	1	1	1	0	0	1	1	1		Unknown	1	1	1	1	1	0	0	1	1	0	0
The Health Improvement Network - Primary care	THIN	United Kingdom	Included		Electronic medical records	Primary care	2002	1	1	1	1	1	1	3	2		3	1	1	1	1	1	1	0	1	1	2	1
QuintilesIMS LPD Health Search Database Longitudinal	LPD France	France	Included		Electronic medical records	Mixed	2001	1	1	1	0	1	1	2	2		3	0	1	0	1	1	0	0	1	1	0	0
Caserta Database	Caserta	Italy	Included		Claims	Primary care	2002	1	0	1	1	1	1	1	1	100%	1	1	1	0	1	1	0	0	1	1	0	0
Medicines Monitoring Unit Scotland	MEMO	Scotland	Included		Record linkage system	Mixed	1990	1	1	1	0	Unknown	1	1	1	20%		1	1	1	1	1	Unknown	Unknown	1	1	Unknown	0
Estonian Health Insurance Fund	EHIF	Estonia	Excluded	Exposure only	Claims	Primary care	2010	1	1	1	0	1	0	1	1	95%	Unknown	1	1	0	1	1	0	1	1	1	0	0

STROBE Statement—Checklist of items that should be included in reports of ***cross-sectional studies***

	Item No	Recommendation	Yes/No/NA	Page No.
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	√	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	√	2
Introduction				
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	√	5
Objectives	3	State specific objectives, including any prespecified hypotheses	√	6
Methods				
Study design	4	Present key elements of study design early in the paper	√	6
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	√	6-7
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants	√	6*
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	NA	
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	√	7
Bias	9	Describe any efforts to address potential sources of bias	NA	
Study size	10	Explain how the study size was arrived at	NA	
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	NA	
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	NA	
		(*b*) Describe any methods used to examine subgroups and interactions	NA	
		(*c*) Explain how missing data were addressed	NA	
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	NA	
		(*e*) Describe any sensitivity analyses	NA	
Results				
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	NA	
		(b) Give reasons for non-participation at each stage	NA	
		(c) Consider use of a flow diagram	√	9 , Figure

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

				2
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	NA	
		(b) Indicate number of participants with missing data for each variable of interest	NA	
Outcome data	15*	Report numbers of outcome events or summary measures	NA	
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	NA	
		(*b*) Report category boundaries when continuous variables were categorized	NA	
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	NA	
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	NA	
Discussion				
Key results	18	Summarise key results with reference to study objectives	√	13-14
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	√	**16**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	√	15-16
Generalisability	21	Discuss the generalisability (external validity) of the study results	NA	
Other information				
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	NA	

* The study unit is study not participants, therefore we answered all questions pertained to patients: selection, counts, as considering study unit.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open

Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-023090.R2
Article Type:	Research
Date Submitted by the Author:	31-Jul-2018
Complete List of Authors:	Pacurariu, Alexandra; European Medicines Agency, Pharmacovigilance and Epidemiology Department Plueschke, Kelly; Pharmacovigilance and Epidemiology McGettigan, Patricia; Queen Mary University of London, William Harvey Research Institute; European Medicines Agency Morales, Daniel; European Medicines Agency (EMA), Surveillance & Epidemiology Service; University of Dundee, Division of Population Health Sciences Slattery, Jim; European Medicines Agency, Surveillance & Epidemiology Service Vogl, Dagmar; European Medicines Agency Goedecke, Thomas; European Medicines Agency Kurz, Xavier; European Medicines Agency, Cave, Alison; European Medicines Agency, Pharmacovigilance and Epidemiology Department
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Health policy
Keywords:	electronic healthcare databases, post-authorisation studies, regulatory science, benefit-risk evaluation, real-world data

SCHOLARONE™
Manuscripts

25th July 2018

Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation

Alexandra Pacurariu*[1,2], Kelly Plueschke[1], Patricia McGettigan[1,3], Daniel R. Morales[1,4], Jim Slattery[1], Dagmar Vogl[1], Thomas Goedecke[1], Xavier Kurz[1], Alison Cave[1]

1 Surveillance & Epidemiology Service, European Medicines Agency (EMA), 30 Churchill Place, London E14 5EU, United Kingdom

2 Dutch Medicines Evaluation Board, 3531 AH Utrecht, the Netherlands

3 William Harvey Research Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK

4 Division of Population Health Sciences, University of Dundee, Dundee, United Kingdom

* Corresponding author: Alexandra Pacurariu alexandra.pacurariu@ema.europa.eu, +44(0)2036608559

Word count:

Number of figures and tables:

3 figures and 2 tables in the manuscript

Key words:

Electronic healthcare databases; post-authorisation studies; regulatory science; benefit-risk evaluation; real-world data; epidemiology.

For peer review only

Abstract

Objective

Electronic healthcare databases (EHDs) are useful tools for drug development and safety evaluation but their heterogeneity of structure, validity and access across Europe complicates the conduct of multi-database studies. In this paper, we provide insight into available EHDs available to support regulatory decisions on medicines.

Methods

EHDs were identified from publicly available information from the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) resources database, textbooks and web-based searches. Databases were selected using criteria related to accessibility, longitudinal dimension, recording of exposure and outcomes, and generalizability. Extracted information was verified with the database owners.

Results

A total of 34 EHDs wwas selected after applying key criteria relevant for regulatory purposes. The most represented regions were Northern, Central and Western Europe. The most frequent types of data source were electronic medical records (44.1%) and record linkage systems (29.4%). The median number of patients registered in the 34 data sources was 5 million (range 0.07-15 million) while the median time covered by a database was 18.5 years. Pediatric patients were included in 32 databases (94%). Completeness of information on drug exposure was variable. Published validation studies were found for only 17 databases (50%). Some level of access exists for 25 databases (73.5%) and 23 databases (67.6%) can be linked through a personal identification number to other databases with parent-child linkage possible in 7 (21%) databases. Eight databases (23.5%) were already transformed or were in the process of being transformed into a common data model that could facilitate multi database studies.

Conclusion

Very few European databases meet minimal regulatory requirements and are readily available to be used in a regulatory context. . Accessibility and validity informationof the included information needs to be improved. . This study confirmed the fragmentation, heterogeneity and lack of transparency existing in many European electronic healthcare databases.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Strengths and limitations of this study

- Data extraction was based on information provided by database owners as well as as on publicly available information
- Incomplete data capture cannot be excluded, especially for very small and new databases.
- Validation of the data source was only evaluated indirectly through the validation studies reported by the database owners
- The inventory was endorsed by an expert working group of the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (the ENCePP Working Group "Data Sources")

For peer review only
DRAFT

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1. Introduction

The European Union (EU) medicines regulatory network has responsibility for protecting patients by ensuring continuous evaluation of the safety of authorised medicines. At the core of such review is the scientific assessment of all available evidence including relevant information from the literature, results from non-clinical studies, randomised clinical trials (RCTs), observational studies, spontaneous reports and results of other available research. A way to collect more information about a medicine's safety post-marketing is by means of post-authorisation safety studies (PASS) (1). PASS may be imposed on a marketing authorisation holder by a regulatory authority or conducted by the company to address a safety concern or evaluate the effectiveness of risk minimisation measures aimed at reducing the occurrence or severity of an adverse reaction (2,3).

Secondary use of routinely collected data from electronic healthcare databases (EHDs) is often used in such studies because it is usually faster and cheaper than primary data collection.

A review of pharmaceutical industry-sponsored studies evaluating the effectiveness of risk minimisation measures submitted to the European Medicines Agency (EMA) for cardiovascular, endocrinology or metabolic drugs authorised between 1995 and 2015 found that EHDs were used in 53% of studies evaluating routine risk minimisation measures and in 31% of studies evaluating additional risk minimisation measures (4). A second review of 189 PASS assessed by the EMA between 2012 and 2015 and registered in the European Union electronic Register of Post-Authorisation Studies (EU PAS Register®) reported that secondary use of routinely collected data was found in 33.3% of PASS and 58% among these leveraged electronic health records (EHRs) (5). A third review of a different set of studies registered in the EU PAS Register® as of December 2016 found that 117 studies (37%) used an existing claims or electronic medical records database (6). A fourth review evaluating studies which measured the impact of regulatory interventions found that claims databases were used in 45% of studies, while EHRs were used in 22% of them, the latter being the most utilised type of data sources for such studies (7). The frequent use of EHDs in observational studies was also reported in a wider context in a review of the abstracts of presentations made at the International Conference for Pharmacoepidemiology: 53% (in 2000) and 51% (in 2005) of submitted EU pharmacoepidemiological studies were conducted using automated general practice, pharmacy or claims data (8).

The fact that between 30%-50% of observational post-authorisation studies use EHDs as their main data source reflects the importance of these data sources to support regulatory decision making (1,9).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

On the other hand, the use of EHDs in pre-authorisation research is currently limited and mostly focused on providing historical control data or understanding the natural history of the disease.

As regulatory decisions based on EHDs may have a considerable impact on public health, the quality of the information, the validity and reproducibility of the derived results require close attention, especially when combining data from several data sources or when the original data is transformed before analysis (10–12). It has been emphasised that the same level of scientific rigour should be employed irrespective of the study design and data source to be used, and that the strengths and weaknesses of each data source should be considered (13). The speed at which the results could be generated is an additional important consideration, particularly for regulatory purpose (9,14,15). By considering the characteristics of the data sources and the research objectives to be addressed, the investigators should be able to choose the most appropriate resource(s) to address the question at hand. However, while some authors provide a detailed description of the databases used in their study (16–19), in other cases thedescription is often incomplete and a justification for their choice in the context of alternative data sources is rarely provided (18). The International Society of Pharmacoepidemiology (ISPE) has developed guidelines to support the selection and use of data sources for observational research by highlighting potential limitations of databases and recommending testing procedures. The guidelines also provide a checklist covering six areas: database selection, use of multiple data resources, extraction and analysis of the study population, privacy and security, quality and validation procedures and documentation (20). The availability of an inventory of European databases describing the main characteristics, conditions of access and validation performed would support investigators to identify databases suitable for their research question. Moreover, knowledge of the characteristics of the data sources used in a post-authorisation study would enhance regulators' confidence in the evidence derived from such data and ultimately in the usefulness of the study in the decision-making process (21,22).

The main objective of this study is to provide an inventory of EHDs and describe their key characteristics and availability with the aim to support stakeholders in their choice of the data source when conducting a post authorisation study.

2. Methods

2.1. Identification of EHDs

As a first step, we identified existing EHDs in Europe by screening the following sources: the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) resources database

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

(21), web-based search engines (22), textbooks on clinical pharmacoepidemiology (23,24), publicly available inventories created for European Commission funded research projects and databases used in EMA funded post-authorisation studies.

As a second step, data sources were included in the inventory based on the following regulatory relevant criteria: the data is available to regulatory authorities or to third-parties for research purposes; the database contains information on both drug exposure and health outcomes and is not disease or product specific; there is longitudinal data capture. Provision of relevant data for benefit-risk decision-making was one of the key criteria for selecting studies meeting regulatory requirements. Prescription-only databases were excluded because they cannot be used for etiological studies in the absence of the outcome recording. Product or disease specific registries were considered out of scope as they create cohorts of patients whose entry is defined either by exposure to a product or by occurrence of a disease or health outcome (25).

Product specific registries are frequently used for the benefit-risk monitoring of specific products, however they rarely cover a wide range of medicines and health conditions and have a narrow scope. Databases where the data collection ceased and the historical data is not accessible were also excluded.

2.2. Data extraction and classification

For each database, publicly available information (on the databases' websites or in publications) was supplemented by contacting data source owners in writing. A total of 82% database owners responded. Teleconferences with seven database owners were conducted to clarify some of the information provided.

The information was extracted by six EMA reviewers (AP, KP, PMG, DM, JS, AC) and the entries for each database were cross-checked for consistency by a second reviewer. Uncertainties about the classification of any variable were resolved through discussion.

The data sources were classified in three categories according to their structure, purpose and type of data: electronic medical records, claims databases and healthcare record linkage systems (e.g. several databases are linked to form a complete database).

The different population registries within the same country were considered as a single national EHD if they could (and are routinely used as such) be linked using a unique identification number (e.g. in Nordic countries and in Scotland).The size of the data source was quantified by the cumulative number

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

of patients included (both total and active patients) and number of years since the initiation of data collection in the database.

Data collected in the following categories was also recorded: demographic information (age and gender of each individual), information on prescribed or dispensed medicines (including name, dose, duration, route of administration and therapeutic indication), immunisations, diagnosis data and referrals for laboratory investigations, imaging, and other procedures. Information of laboratory tests results was not collected.

2.3. Availability of validation studies

Database owners were asked to report validation studies which they were aware of for their database. Studies published up to September 2016 were included. For the purpose of this study, a validation study was defined as any study published in a peer-reviewed journal that aimed to validate the information available on an outcome or exposure in comparison with gold standard information, usually the patients' original health records as reviewed by a medical professional or the same information captured by another database for a different purpose. For example, a study in 2012 compared cancer records in a general practitioners' database, hospital records and cancer registries and found considerable discrepancies in cancer recording between these different data sources (26).

2.4. Accessibility

The accessibility of databases for research purposes was classified in four categories: no access, indirect access through the database owner or a third party, direct access restricted to specific datasets and direct access to the full dataset.

2.5. Coding of database characteristics for usefulness for medicines benefit-risk evaluation

Instead of evaluating the quality of each database, we aimed to assist in the selection of databases by implementing a coding process that identifies the data sources considered to provide sufficient information to contribute to regulatory questions on the benefit-risk evaluation of medicines. For this purpose, the following domains were included in the coding process: extent of data capture of study variables, size of data source, quality and validity of information, accessibility, potential for linkage and existing process in place to convert the data into a common data model (CDM) (Figure 1). A common data model provides a common representation and architecture of the data across multiple databases,

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

thus enabling the standardisation of administrative and clinical information and allowing the use of common analytical tools (27).

[Insert Figure 1. Coding of the characteristics of electronic healthcare databases available in Europe for the benefit-risk evaluation of medicines.]

The ENCePP Working Group "Data Sources" (9) reviewed an initial version of the inventory with the description of databases and endorsed the final inventory.

Patient involvement statement: This descriptive analysis did not involve any patients.

3. Results

3.1. General overview

The initial search generated a list of 77 potential data sources. After merging the national registries into a single entry and applying the exclusion criteria, 34 of them were retained in the final inventory (Figure 2). Table 1 provides a list of these 34 databases and the complete information is provided in the *Supplementary material.*

[Insert Figure 2. Flowchart of database selection]

For peer review only

DRAFT

Table 1. List of data sources retained in the final inventory (by year)

Data source name	Country	Type	Type of care	Start date
Finnish National registries	Finland	Record linkage system	Mixed	1964
Swedish National registries	Sweden	Record linkage system	Mixed	1970
Danish National and regional registries	Denmark	Record linkage system	Mixed	1977
The electronic Data Research and Innovation Service	Scotland	Record linkage system	Mixed	1981
Clinical Practice Research Datalink	United Kingdom	Electronic medical records	Primary care	1987
QResearch	United Kingdom	Electronic medical records	Primary care	1989
Information System of Parc de Salut del Mar	Spain	Electronic medical records	Primary care	1990
Medicines Monitoring Unit Scotland	Scotland	Record linkage system	Mixed	1990
Pharmo Database Network	Netherlands	Record linkage system	Mixed	1990
QuintilesIMS Disease Analyser	Germany	Electronic medical records	Mixed	1992
Integrated Primary Care Information Database	Netherlands	Electronic medical records	Primary care	1995
Agencia Regionale di Sanita Tuscany database	Italy	Claims	Secondary care	1996
Norwegian Registries	Norway	Record linkage system	Mixed	1997
QuintilesIMS Disease Analyser	France	Electronic medical records	Primary care	1997
Region Emilia Romagna Database	Italy	Claims	Secondary care	1997
Hospital Information System -Lazio	Italy	Claims	Secondary care	1998
Icelandic Registries	Iceland	Record linkage system	Mixed	1998
Pedianet Database	Italy	Electronic medical records	Primary care	1998
Securite Sociale de l'Assurance Maladie	France	Claims	Mixed	1999
Lombardia Health Database	Italy	Electronic medical records	Secondary care	2000
QuintilesIMS LPD Health Search Database Longitudinal	Italy	Electronic medical records	Primary care	2000
QuintilesIMS LifeLink: Hospital Disease Database - Belgium	Belgium	Electronic medical records	Secondary care	2001
QuintilesIMS LPD Health Search Database Longitudinal	France	Electronic medical records	Mixed	2001
Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria	Spain	Electronic medical records	Primary care	2002
Caserta Database	Italy	Claims	Primary care	2002
The Health Improvement Network	United Kingdom	Electronic medical records	Primary care	2002
German Pharmacoepidemiological Research Database	Germany	Claims	Mixed	2004
Echantillon Généraliste de Bénéficiaires	France	Claims	Mixed	2006
QuintilesIMS LPD Health Search Database Longitudinal	Spain	Electronic medical records	Mixed	2006
The Information System for the Development of Research in Primary Care	Spain	Electronic medical records	Primary care	2006
VEKTIS	Netherlan	Claims	Mixed	2006

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	ds			
Secure Anonymised Information Linkage	Wales	Record linkage system	Mixed	2007
National Health Fund	Poland	Claims	Mixed	2008
Hospital Treatment Insights	United Kingdom	Record linkage system	Secondary care	2010

The most frequent types of data source identified were electronic medical records (n=15, 44.1%) followed by record linkage systems (n=10, 29.4%) and claims databases (n=9, 26.5%). In terms of the type of care covered, mixed care settings (primary and secondary care) were most common (n=17, 50%), followed by primary care databases (n=11, 32.3%) (Table 2). The median number of patients followed cumulatively across the 34 data sources was 5 million (range 0.07-15 million).

Table 2. Distribution of data sources type and type of care covered

Type of data source	Primary care	Secondary care	Mixed
Claims	1 (2.9%)	3 (8.82%)	5 (14.7%)
Electronic medical records	10 (29.4%)	2 (5.9%)	3 (8.8%)
Record linkage system	0 (0%)	1 (2.9%)	9 (26.5%)

Patient age and gender were recorded in all data sources while paediatric patients were included in 32 databases (94%). The median year for database start was 1998, with the oldest database established in 1964 (the Finnish Hospital Discharge Register). The median calendar time covered by a database was 18.5 years (range 7-53 years). In terms of geographical coverage, 17% of databases collect data from Norway, 14% from Finland and 10% from Denmark and Italy (Figure 3).

[Insert
Figure 3: European data sources and duration of data collection.

3.2. Information captured

By definition, all the databases retained in the final inventory contained information about drug exposure (either prescribed or dispensed). The completeness of information was however variable: 28 databases (82.3%) had information about prescribed dose and duration of treatment (either directly recorded or inferred from other collected variables); 14 (41.1%) had information about route of administration; 20 databases (58.8%) recorded the therapeutic indication associated with the prescription (either directly recorded or inferred from other database elements). Over-the-counter

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

drugs were rarely and inconsistently captured in any of the databases while vaccinations were captured in 13 databases (38.0%). Data on hospital in-patient administered drugs were rarely captured (5.8%). All databases had information about medical events (diagnosis) as a prerequisite for inclusion in our inventory. Referrals for laboratory investigations were captured in 19 (55.9%) and referrals for imaging or other diagnostic procedures were captured in 16 databases (47.1%).

3.3. Validation studies

No published validation study was reported for 17 databases (50.0%), while a total of 42 validation studies were reported for the other 17 databases, with a median of 3 validation studies per database (range: 1-25). The validation concerned either specific health outcomes or prescription information. The most common gold standards used for the validation included paper based prescriptions, medical records, death records and perinatal deaths obtained from registries or national statistics reports. Some database owners have reported as validation studies the validation of prediction algorithms for various health outcomes as chronic kidney disease, ischaemic stroke and various types of cancers based on an estimating the absolute risk of a particular outcome in primary care patients with and without symptoms (1,2). It is debatable if these are truly validation studies according to our definition.

3.4. Accessibility and potential for linkage

One database was excluded due to lack of access to third parties, 10 (29.4%) offer indirect access to the database for third parties, 7 (20.5%) provide direct access to specific datasets and 8 (23.5%) offer direct access to the full dataset. The level of access could not be identified for 8 EHDs (23.5%). In terms of linkage, 23 databases (67.6%) could be linked through a unique personal identification number (PIN) to other databases containing additional healthcare-related information including cause of death registries, hospital data, prescription databases and cancer registries. The Nordic registries are a good example of extensive linkage among different national registries through usage of a PIN. Other forms of linkage do exist, for example, in order to avoid the use of PIN and preserve anonymity, the PHARMO network uses probabilistic linkage based on patient birth date, gender and general practitioner code. The linkage of a parent with their child ('parent-child linkage'), which is useful for studies investigating pregnancy exposures and effect on offspring, was available in 7 data sources (20.6%).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

3.5. Conversion of the database to a CDM

Four (11.7%) databases were already transformed in a CDM and four others were in the process of being converted to a CDM (the QuintilesIMS Disease Analyser France and Germany, the Spanish Information System for the Development of Research in Primary Care, the Agenzia Regionale di Sanità Tuscany database, The Pedianet, the Clinical Practice Research Datalink, the Integrated Primary Care Information Database and The Health Improvement Network). Seven of these 8 databases used the Observational Medical Outcomes Partnership CDM (27), while the Spanish Information System for the Development of Research in Primary Care (28) is implementing the model used in the ADVANCE project for vaccine studies (29).

4. Discussion

A total of 34 European EHDs with potential for use in the regulatory environment were included in this study. The most frequently represented regions s were Northern, Central and Western Europe, with a scarcity of data sources in Eastern Europe. The most common data sources assessed were electronic medical records with a mix of primary and secondary care coverage. Most of the databases contain out-patient prescribing while in-patient prescribing is very rarely captured. The median number of patients registered within the 34 data sources was 5 million and the median calendar time covered by a database was 18.5 years. In terms of accessibility, 24% of databases offered direct access to the full data source, with the rest having a somewhat more limited access. There are a few similar studies of EHDs available in Europe (8,30), but as far as we are aware this is the first study taking a regulatory perspective. An analysis of the characteristics of post-authorisation studies requested by regulators showed that 47% of studies involved secondary use of data emphasizing the important role of secondary data in the regulatory setting. More detailed descriptions of database characteristics are provided in electronic repositories such as the European Medical Information Network (EMIF), the ENCePP resource database and the Bridge to Data initiative (21,22). However existing repositories are either incomplete, have a limited coverage or they require a fee for access, therefore restricting access to their information.

This study helps identify databases with key characteristics as an entry door to further investigate with their owner their potential usefulness for a specific study.

Given that different national guidelines and clinical practice can generate significant heterogeneity in how healthcare is delivered and recorded (31), it is important that regulators have access to data from

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

as broad a geographical spread as possible. Thus there is a clear need for the development of data sources in EU member states which currently either have no data sources or are poorly represented. The data recorded in the databases include some limitations. Firstly, the limited capture of in-patient prescribing poses a problem for regulators and investigators since many newly approved drugs are specialised drugs, used exclusively in secondary care (32). Secondly, some disease-specific variables (e.g., biomarkers, laboratory tests and genetic data) are only exceptionally recorded and they are required more and more often in study protocols. High quality disease registries can to some extent meet this need in specific disease areas but they rarely capture co-medications, co-morbidities and adverse reactions. Improvements in the quality of in-patient care and in the recording of laboratory tests would be of value for epidemiological investigations on determinants for health outcomes, including drug-related safety issues.

With regards to validation, 50% of databases had at least one validation study published. Validation should normally be done for the data elements collected in every study. Publication of validation studies is not an indicator of the overall validity of the database but may inform researchers on the feasibility to perform study-specific validation in a database. A repository of validated outcomes in specific databases would reduce duplication of work. Such a repository should include a clear description of the methodology and limitations of the analysis.

Extending approved adult indications to the paediatric population is increasing and according to the European Commission's report between 50% and 90% of the medicines currently used in pediatrics have neither been tested on nor authorised for use in children (32). Availability of real-world data is therefore particularly important for this purpose. In our review we found that 94% of databases have some information about pediatric patients but no in-depth analysis of the available information was undertaken. A more detailed review of pediatric databases was undertaken by Neubert et al. who concluded that in Europe, drug utilisation and outcome data is available for ~4 million children (33). However, similar to our study, the authors highlight that efforts should be made to increase availability of in-patient data, a setting where the greatest prescribing of novel medicines occurs (33).

While validity studies were published for half of the databases, van Staa *et al*. (15) highlighted that systematic measurement of data quality is lacking in most databases. As such and in line with the recommendations of Hall et al (20), we encourage data holders to document the basic characteristics of their data source and to highlight when a change in recording practices occurs.

A new way forward to increase the speed and power of multi-centres studies is the use of a CDM (34). The advantage of using a CDM is that the transformed databases can be more easily integrated for

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

research across a network. Although less than a third of databases were already converted or in process of being converted to a CDM in Europe, these figures are likely to change fast due to ongoing initiatives such as EMIF (35) and the European Health Data Network project (EHDN) (36).

Access to databases for research purposes can be provided at patient level in only a quarter of databases while the remaining ones had more restrictive access policies. We therefore fully support the recommendations published by other groups that governance models should be in place to facilitate data access, data sharing and secondary use of research data in health sciences (37).

There are multiple challenges to the utilisation of EHDs in a regulatory context, particularly in Europe, which go beyond the above mentioned challenges related to the characteristics of the specific databases. These include fragmentation and lack of interoperability of European data sources, inconsistent use of methods to integrate and analyse heterogeneous data, lack of systematic and consistent validation of data sources, governance issues and privacy concerns. In an attempt to deal with the significant heterogeneity across data sources in Europe, ENCePP has established a Working Group dedicated to facilitating the initiation and conduct of observational research using multiple data sources (9). As part of its work the group reviewed ongoing or finalised multi-database drug safety projects of various publicly funded EU projects, which highlighted the heterogeneity of the methods used for combining EHR data from multiple databases (3). Ongoing work of the group is centred around developing guidance on conceptual models for multi-national and multi-database studies.

Our review has a number of limitations. Firstly, we may have missed data sources during the identification process. However, we attempted to be as complete as possible by incorporating several rounds of database identification and review of the inventory by experts, including members of the ENCePP Working Group "Data sources" and database owners. The difficulties we encountered when trying to map all the existing EHDs in Europe highlight again the need for more comprehensive and accessible repositories with EHDs.

Secondly, we excluded prescription only databases since they cannot be used for etiological studies even if we acknowledge their utility for drug utilisation studies which are very common in the regulatory field. Lastly, validation of the primary source data is an important process that provides confidence in the results of the analyses (38) and this was only evaluated indirectly through the number of validation studies reported by the database owners. A strength of our study was that data from publicly available sources was complemented or verified with database owners.

There is more work to be done in order to increase transparency and accessibility of existing datasources. Examples of areas for future development are to develop more robust validation

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

measures and increase transparency of validated outcomes, to transform databases through a CDM to allow faster feasibility assessment and execution of studies, and to stimulate creation and access to EHRs in Eastern Europe.

5. Conclusion

We have provided a systematic inventory of EHDs available in Europe that includes a summary evaluation of their capability to support regulatory decision-making on the benefits and risks of medicines in Europe. Despite the wide range of health care databases available for epidemiologic research in Europe, many of them were excluded from the inventory due to the absence of information needed for key regulatory activities. The analysis of the included databases confirmed the fragmentation, heterogeneity and lack of transparency existing in European electronic healthcare databases.

The analysis has focussed on population–based electronic healthcare databases allowing conducting causal association studies between drug exposure and health outcomes in primary care. Our intention is to help the identification of and access to relevant existing databases that could be used for public health research. Beyond this objective, we consider that this inventory may assist clinical epidemiologists interested in undertaking other investigations such as studying the occurrence and determinants of health outcomes in a population.

We hope that this inventory should stimulate increased transparency and accessibility of other databases in addition to the development of data sources in Eastern European countries, which are currently under-represented.

For peer review only
DRAFT

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

References

1. Good pharmacovigilance practice. Module VIII. Post-authorisation safety studies. 2016; EMA/813938/2011 Rev 3* [Accessed 2017 Nov 17]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129137.pdf

2. Santoro A, Genov G, Spooner A, Raine J, Arlett P. Promoting and Protecting Public Health: How the European Union Pharmacovigilance System Works. Drug Saf. 2017 Oct;40(10):855–69; DOI: 10.1007/s40264-017-0572-8.

3. Blake KV, Prilla S, Accadebled S, et al. European Medicines Agency review of post-authorisation studies with implications for the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. Pharmacoepidemiol Drug Saf. 2011 Oct;20(10):1021–9.

4. Mazzaglia G, Straus SMJ, Arlett P, et al. Study Design and Evaluation of Risk Minimization Measures: A Review of Studies Submitted to the European Medicines Agency for Cardiovascular, Endocrinology, and Metabolic Drugs. Drug Saf. 2017 Nov 9; DOI: 10.1007/s40264-017-0604-4.

5. Engel P, Almas MF, De Bruin ML, Starzyk K, Blackburn S, Dreyer NA. Lessons learned on the design and the conduct of Post-Authorization Safety Studies: review of 3 years of PRAC oversight. Br J Clin Pharmacol. 2017 Apr;83(4):884–93; DOI: 10.1111/bcp.13165.

6. Carroll R, Ramagopalan SV, Cid-Ruzafa J, Lambrelli D, McDonald L. An analysis of characteristics of post-authorisation studies registered on the ENCePP EU PAS Register. F1000Research. 2017 Nov 10;6:1447; DOI: 10.12688/f1000research.12198.2.

7. Goedecke T, Morales DR, Pacurariu A, Kurz X. Measuring the impact of medicines regulatory interventions - Systematic review and methodological considerations. Br J Clin Pharmacol. 2017 Dec 20; doi: 10.1111/bcp.13469.

8. Sturkenboom MCJM. Other Databases in Europe for the Analytic Evaluation of Drug Effects. Pharmacovigilance. John Wiley & Sons, Ltd; 2007. p. 361–73; DOI: 10.1002/9780470059210.ch29.

9. Kurz X, Perez-Gutthann S, ENCePP Steering Group. Strengthening standards, transparency, and collaboration to support medicine evaluation: Ten years of the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Pharmacoepidemiol Drug Saf. 2018 Jan 11; DOI: 10.1002/pds.4381

10. Klungel OH, Kurz X, de Groot MCH, et al. Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project. Pharmacoepidemiol Drug Saf. 2016 Mar;25 Suppl 1:156–65; DOI: 10.1002/pds.3968

11. Trifirò G, Coloma PM, Rijnbeek PR et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? J Intern Med. 2014 Jun 1;275(6):551–61; DOI: 10.1111/joim.12159

12. La Gamba F, Corrao G, Romio S, et al. Combining evidence from multiple electronic health care databases: performances of one-stage and two-stage meta-analysis in matched case-control studies. Pharmacoepidemiol Drug Saf. 2017 Oct;26(10):1213–9; DOI: 10.1002/pds.4280

13. Strom B. How Should One Perform Pharmacoepidemiology Studies? Choosing Among the Available Alternatives - Pharmacoepidemiology, Third Edition, Wiley Online Library, 2013, p.401-413; DOI: 10.1002/9781118344828.ch11

14. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005 Apr;58(4):323–37; DOI: 10.1016/j.jclinepi.2004.10.012

15. van Staa TP, Klungel OH. Real-life data and learning from practice to advance innovation. Available from: http://www.who.int/medicines/areas/priority_medicines/BP8_4Data.pdf

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

16. Abbing-Karahagopian V, Kurz X, de Vries F, et al. Bridging differences in outcomes of pharmacoepidemiological studies: design and first results of the PROTECT project. Curr Clin Pharmacol. 2014 May;9(2):130–8; DOI: 10.2174/1574884708666131111211802

17. Coloma PM, Schuemie MJ, Trifirò G, 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; DOI: 10.1002/pds.2053

18. Madigan D, Ryan PB, Schuemie M, et al. Evaluating the impact of database heterogeneity on observational study results. Am J Epidemiol. 2013 Aug 15;178(4):645–51; DOI: 10.1093/aje/kwt010

19. Roberto G, Leal I, Sattar N, et al. Identifying Cases of Type 2 Diabetes in Heterogeneous Data Sources: Strategy from the EMIF Project. Virgili G, editor. PLOS ONE. 2016 Aug 31;11(8):e0160648; DOI: 10.1371/journal.pone.0160648

20. Hall GC, Sauer B, Bourke A, et al. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiol Drug Saf. 2012 Jan;21(1):1–10; DOI: 10.1002/pds.2229

21. ENCePP Resources Database [Internet]. [Accessed 2018 Jan 1]. Available from: http://www.encepp.eu/encepp

22. Bridge to data [Internet]. [Accessed 2018 Jan 1]. Available from: https://www.bridgetodata.org/

23. Strom BL, editor. Pharmacoepidemiology Chichester, UK: John Wiley & Sons, Ltd; 2000.]. [Accessed 2018 Jan 1]. Available from: http://doi.wiley.com/10.1002/0470842555; DOI: 10.1002/9781118344828

24. Rothman KJ. Epidemiology: An Introduction. Oxford University Press, USA; 2002

25. PARENT Joint Action. Methodological guidelines and recommendations for efficient and rational governance of patient registries. 2015. [Accessed 2018 Jan 1]. Available from: http://patientregistries.eu/deliverables

26. Boggon R, van Staa TP, Chapman M, Gallagher AM, Hammad TA, Richards MA. Cancer recording and mortality in the General Practice Research Database and linked cancer registries. Pharmacoepidemiol Drug Saf. 2013 Feb;22(2):168–75; DOI: 10.1002/pds.3374

27. Gagne JJ. Common Models, Different Approaches. Drug Saf. 2015 Aug;38(8):683–6.

28. Garcia-Gil M, Hermosilla E, Prieto-Alhambra D, Fina F, Rosell M, et al. (2011) Construction and validation of a scoring system for selection of high quality data in a Spanish population primary care database (SIDIAP). Inf Prim Care.19(3): 135–45.

29. ADVANCE Work Package 5 White Paper Proof-of-concept studies [Accessed 2018 May 23]. Available from : http://www.advancevaccines.eu/?page=publications&id=DELIVERABLES

30. Wettermark B, Zoëga H, Furu K, Korhonen M, Hallas J, Nørgaard M, et al. The Nordic prescription databases as a resource for pharmacoepidemiological research--a literature review. Pharmacoepidemiol Drug Saf. 2013 Jul;22(7):691–9; DOI: 10.1002/pds.3457

31. Health at a Glance: Europe 2016 | OECD READ edition [Internet]. OECD iLibrary. [Accessed 2018 Jan 15]. Available from: http://www.keepeek.com/Digital-Asset-Management/oecd/social-issues-migration-health/health-at-a-glance-europe-2016_9789264265592-en

32. EMA Annual Report 2016 [Internet]. [Accessed 2018 Jan 1] Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Annual_report/2017/05/WC500227334.pdf

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

33. Neubert A, Sturkenboom MCJM, Murray ML, Verhamme KMC, Nicolosi A, Giaquinto C, et al. Databases for pediatric medicine research in Europe-assessment and critical appraisal. Pharmacoepidemiol Drug Saf. 2008 Dec;17(12):1155–67; DOI: 10.1002/pds.1661

34. Gini R, Schuemie M, Brown J, et al. Data Extraction and Management in Networks of Observational Health Care Databases for Scientific Research: A Comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies. EGEMS Wash DC. 2016;4(1):1189; DOI: 10.13063/2327-9214.1189

35. European Medical Information Framework [Internet]. [Accessed 2018 Jan 1]. Available from: http://www.emif.eu/

36. European Health Data Network. 2018. [Internet]. [Accessed 2018 Jan 1]. Available from: http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/imi2-2017-12-04.html

37. Burton PR, Banner N, Elliot MJ, Knoppers BM, Banks J. Policies and strategies to facilitate secondary use of research data in the health sciences. Int J Epidemiol. 2017 Dec 1;46(6):1729–33; DOI: 10.1093/ije/dyx195

38. Ehrenstein V, Petersen I, Smeeth L, et al. Helping everyone do better: a call for validation studies of routinely recorded health data. Clin Epidemiol. 2016;8:49–51; DOI: 10.2147/CLEP.S104448

List of figures' and table's titles

Figure 1. Coding of the characteristics of electronic healthcare databases available in Europe for the benefit-risk evaluation of medicines.

Figure 2. Flowchart of database selection

Figure 3. European data sources and duration of data collection

Table 1. List of data sources retained in the final inventory (by year)

Table 2. Distribution of data sources type and type of care covered

Footnotes

Contributions: All authors (AP, KP,PMG, DM, JS, DV, TG, XK and AC) were involved in the study design and data collection. AP, XK, and DM performed the analysis and interpretation of results. AP, DM and AC contributed to writing and KP,PMG, JS, TG, XK revised and approved the final draft.

Funding: No sources of funding were used to assist in the preparation of this article.

Copyright statement: The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non-exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd to permit this article (if accepted) to be published in BMJ editions and any other BMJPGL products and sublicenses such use and exploit all subsidiary rights, as set out in our licence.

Conflict of interest: We have read and understood the BMJ Group policy on declaration of interests and declare the following interests: None

Disclaimer: The views expressed in this article are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the European Medicines Agency or one of its committees or working parties.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Ethical approval: Not required.

Data sharing: An extended version of the dataset is available as supplementary material.

Transparency: AP affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Open Access: This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

DRAFT
For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Figure 1. Coding of European electronic healthcare databases characteristics for the benefit-risk evaluation of medicines.
The coding system was binary: 0 if information was absent and 1 if it was present. The degree of completion for a specific variable was not recorded. An exception to the binary classification was done for the accessibility variable: 0 - no access; 1 - indirect access through database owner or third party; 2 - direct access to specific data sources; 3 - direct access to full data source.

150x150mm (300 x 300 DPI)

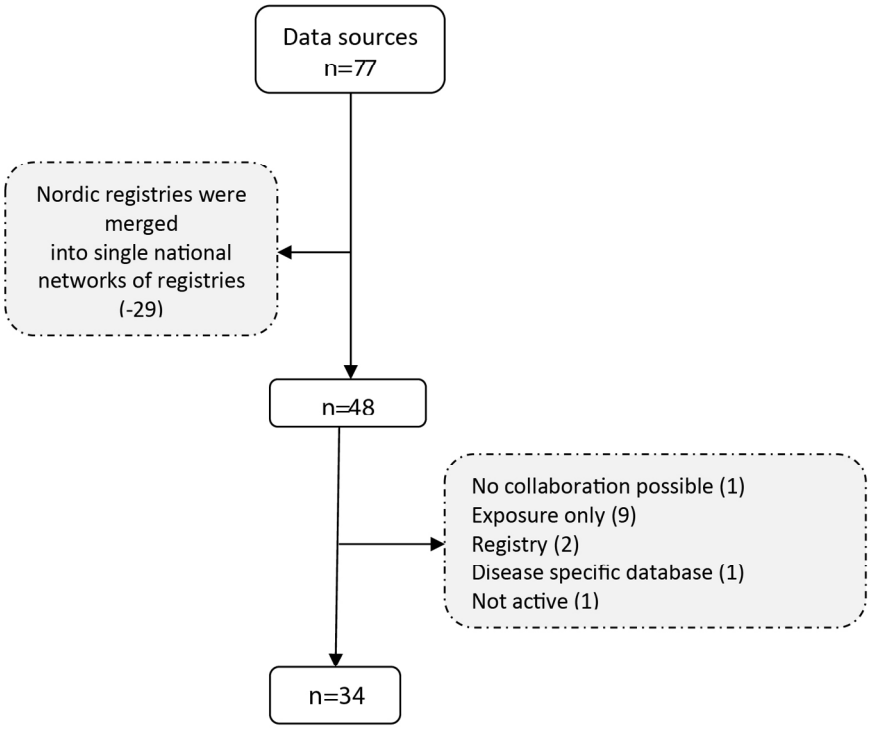


Figure 2. Flowchart of database selection

150x150mm (300 x 300 DPI)

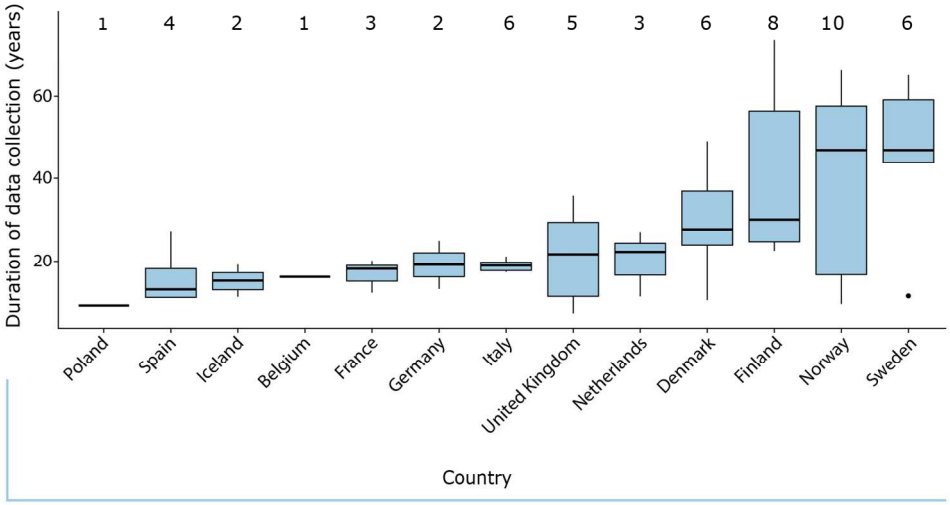


Figure 3. European data sources and duration of data collection
Boxplots indicate the median (horizontal black line) data collection time by country while the margins of the boxplot represent the interquartile range, the vertical lines indicate the minimum and maximum values. The number of databases per country are provided above the boxplots.

150x150mm (300 x 300 DPI)

Data source name	Abbreviation	Country	Included/Excluded	Exclusion criteria	Type	Type of care	Start date	Prescribed medicines	Dose	Duration	Route of administration	Indications	Diagnosis	Number of patients	Number of active patients	Coverage	Access and analysis of data	Linkage	Paediatric data	Hospital data	Age and/or date of birth of patients	Gender of patients	Parent-child linkage	Inclusion of vaccinations	Screening results eg scans etc	Test results (eg blood tests, blood pressure, etc)	Studies available	Is data transformed or being transformed to CDM
								0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	1 - (<5M);	1 - (<5M);		0 - No access 1 - Indirect access through atabase	0 - No 1 - Yes	0 - No 1 - Yes	0 - No 1 - Yes	0 - if no assessment of	0 - if no assessment of	0 - if no assessment	0 - if no assessment of	0 - if no assessment of	0 - if no assessment of	0 - if no studies	0 - No 1 - Yes
QuintilesIMS LifeLink: Hospital Disease Database - Belgium	HDD	Belgium	Included		Electronic healthcare	Secondary care	2001	1	1	1	0	1	0	1	1	25%	3	0	1	0	1	1	0	0	1	0	0	0
UK-CRIS- Uk Case Record Interactive search	UK Cris	United Kingdom	Excluded	Disease specific database																								
IMS LifeLink:Longitudinal Prescription Data (LRx) - Belgium		Belgium	Excluded	Exposure only												30%												
Danish National and regional registries		Denmark	Included		Record linkage system	Mixed	1977	1	1	1	1	1	1	2	1	100%	1	1	1	1	1	1	0	1	1	1	2	0
Finnish National registries		Finland	Included		Record linkage system	Mixed	1964	1	1	1	1	0	1	3	3	100%	2	1	1	1	1	1	1	1	0	0	2	0
Longitudinal Prescription Data France	IMS LifeLink Treatment Dynamics	France	Excluded	Exposure only												33%												
Securite Sociale de l'Assurance Maladie	SNIIRAM	France	Included		Claims	Mixed	1999	1	1	1	0	0	1	3	3	90%	3	1	1	0	1	1	0	1	1	1	0	0
QuintilesIMS Disease Analyser	DA France	France	Included		Electronic medical records	Primary care	1997	1	1	1	0	1	1	2	1	2%	3	0	1	0	1	1	0	0	1	1	2	1
Echantillon Généraliste de Bénéficiaires	EGB	France	Included		Claims	Mixed	2006	1	1	1	0	0	1	1	1	1.03%	3	1	1	0	1	1	0	1	1	1	0	0
QuintilesIMS Disease Analyser	DA Germany	Germany	Included		Electronic medical records	Mixed	1992	1	1	1	0	1	1	3	2	3.20%	3	0	1	0	1	1	0	1	1	1	2	1
German Pharmacoepidemiological Research Database	GePaRD	Germany	Included		Claims	Mixed	2004	1	1	1	1	0	1	3	2	17%	1	1	1	1	1	1	1	1	0	0	1	0
Icelandic Registries		Iceland	Included		Record linkage system	Mixed	1998	1	1	1	1	1	1	1	1	100%	3	1	1	1	1	1	0	0	0	0	0	0
Health Service Executive - Primary Care Reimbursement Services	HSE	Ireland	Excluded	Exposure only												100%												
Pedianet Database	PediaNet	Italy	Included		Electronic medical records	Primary care	1998	1	1	1	1	1	1	1	1	3%	1	1	1	1	1	1	0	1	1	1	2	1
Agencia Regionale di Sanita Tuscany database	ARS	Italy	Included		Claims	Secondary care	1996	1	0	0	0	1	1	1	1	100%	Unknown	0	1	1	1	1	0	1	1	1	1	1
Hospital Information System -Lazio	HIS	Italy	Included		Claims	Secondary care	1998	0	0	0	0	1	0	1	1		1	1	1	1	1	1	0	0	0	0	0	0
Lombardia Health Database		Italy	Included		Electronic medical records	Secondary care	2000	1	0	0	0	1	1	2	1	100%	Unknown	1	1	1	1	1	0	0	0	0	0	0
QuintilesIMS LPD Health Search Database Longitudinal	LPD Italy	Italy	Included		Electronic medical records	Primary care	2000	1	1	1	1	1	1	1	1		3	0	0	0	1	1	0	Unknown	Unknown	1	1	0
Region Emilia Romagna Database	RER	Italy	Included		Claims	Secondary care	1997	1	Unknown	Unknown	Unknown	Unknown	1	1	1		Unknown	1	Unknown	1	1	1	0	Unknown	Unknown	Unknown	0	0
OSMED National Drug Consumption Database	OSMED	Italy	Excluded	Exposure only																								
National Health Insurance Fund Database		Lithuania	Excluded	Exposure only												40%												
IMS LifeLink - Longitudinal Prescription Data Netherlands	MS Xtrend Dynamics	Netherlands	Excluded	Exposure only												75%												
Integrated Primary Care Information Database	IPCI	Netherlands	Included		Electronic medical records	Primary care	1995	1	1	1	0	1	1	1	1		1	0	1	1	1	1	0	1	1	1	1	1
VEKTIS	VEKTIS	Netherlands	Included		Claims	Mixed	2006	1	Unknown	Unknown	Unknown	0	1	3	2	100%	2	1	1	Unknown	1	1	0	Unknown	Unknown	Unknown	Unknown	0
Pharmo Database Network	PHARMO	Netherlands	Included		Record linkage system	Mixed	1990	1	1	1	1		1	1	1	25%	1	1	1	0	1	1	0	1	0	0	2	0
Norwegian Registries		Norway	Included		Record linkage system	Mixed	1997	1	1	1	0	1	1	2	2		2	1	1	1	1	1	0	0	0	0	1	0
National Health Fund		Poland	Included		Claims	Mixed	2008	1	1	1	1	0	1	3	3	86%	1	0	1	1	1	1	0	0	0	0	0	0
Programme médicalisé des systèmes d'informations		France	Excluded	No collaboration																								
National Drug Consumption Database (Out-patient Prescribing of Drugs)		Slovenia	Excluded	Exposure only																								
Spanish National Drug Consumption Database		Spain	Excluded	Exposure only												74%												
Osservatorio Interaziendale per la Farmacoepidemiologia e la Farmacoeconomia		Italy	Excluded	Not active																								
The Information System for the Development of Research in Primary Care	SIDIAP	Spain	Included		Electronic medical records	Primary care	2006	1	1	1	0	1	1	2	2		2	1	1	1	1	1	1	1	1	1	2	1
Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria	BIFAP	Spain	Included		Electronic medical records	Primary care	2002	1	1	1	1	1	1	2	1	17%	1	0	1	0	1	1	0	0	0	1	0	0
Information System of Parc de Salut del Mar	IMASIS	Spain	Included		Electronic medical records	Primary care	1990	1	1	1	0		1	1	1		Unknown	1	1	0	1	1	0	0	0	1	0	0
JRC EUROCAT Central Registry		Europe	Excluded	Registry																								
QuintilesIMS LPD Health Search Database Longitudinal	LPD Spain	Spain	Included		Electronic medical records	Mixed	2006	1	1	1	0	1	1	1	1		3	0	1	0	1	1	0	0	0	1	0	0
Swedish National Registries		Sweden	Included		Record linkage system	Mixed	1970	1	1	1	0	0	1	1	1	98%	2	1	1	1	1	1	0	0	0	0	1	0
Clinical Practice Research Datalink - Primary care	CPRD	United Kingdom	Included		Electronic medical records	Primary care	1987	1	1	1	1	0	1	3	2		3	1	1	1	1	1	1	1	1	1	2	1
QResearch	Qresearch	United Kingdom	Included		Electronic medical records	Primary care	1989		1	1	1	1	1	3	3	43%	1	1	1	1	1	1			1	1	2	0
The electronic Data Research and Innovation Service	eDRIS-ISD	Scotland	Included		Record linkage system	Mixed	1981	1	1	1	1	0	1	2	2	100%	2	1	1	1	1	1	1	1	0	0	2	0
Secure Anonymised Information Linkage	SAIL	Wales	Included		Record linkage system	Mixed	2007	1	1	1	Unknown	1	1	3	3	100%	1	1	1	1	1	1	1	Unknown	Unknown	Unknown	0	0
EUROmediCAT	EUROmediCAT	Europe	Excluded	Registry																								
Hospital Treatment Insights	HTI	United Kingdom	Included		Record linkage system	Secondary care	2010	1	1	1	0	0	1	1	1		Unknown	1	1	1	1	1	0	0	1	1	0	0
The Health Improvement Network - Primary care	THIN	United Kingdom	Included		Electronic medical records	Primary care	2002	1	1	1	1	1	1	3	2		3	1	1	1	1	1	1	0	1	1	2	1
QuintilesIMS LPD Health Search Database Longitudinal	LPD France	France	Included		Electronic medical records	Mixed	2001	1	1	1	0	1	1	2	2		3	0	1	0	1	1	0	0	1	1	0	0
Caserta Database	Caserta	Italy	Included		Claims	Primary care	2002	1	0	1	1	1	1	1	1	100%	1	1	1	0	1	1	0	0	1	1	0	0
Medicines Monitoring Unit Scotland	MEMO	Scotland	Included		Record linkage system	Mixed	1990	1	1	1	0	Unknown	1	1	1	20%		1	1	1	1	1	Unknown	Unknown	1	1	Unknown	0
Estonian Health Insurance Fund	EHIF	Estonia	Excluded	Exposure only	Claims	Primary care	2010	1	1	1	0	1	0	1	1	95%	Unknown	1	1	0	1	1	0	1	1	1	0	0

STROBE Statement—Checklist of items that should be included in reports of ***cross-sectional studies***

	Item No	Recommendation	Yes/No/NA	Page No.
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	√	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	√	2
Introduction				
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	√	5
Objectives	3	State specific objectives, including any prespecified hypotheses	√	6
Methods				
Study design	4	Present key elements of study design early in the paper	√	6
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	√	6-7
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants	√	6*
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	NA	
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	√	7
Bias	9	Describe any efforts to address potential sources of bias	NA	
Study size	10	Explain how the study size was arrived at	NA	
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	NA	
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	NA	
		(*b*) Describe any methods used to examine subgroups and interactions	NA	
		(*c*) Explain how missing data were addressed	NA	
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	NA	
		(*e*) Describe any sensitivity analyses	NA	
Results				
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	NA	
		(b) Give reasons for non-participation at each stage	NA	
		(c) Consider use of a flow diagram	√	9 , Figure

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

				2
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	NA	
		(b) Indicate number of participants with missing data for each variable of interest	NA	
Outcome data	15*	Report numbers of outcome events or summary measures	NA	
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	NA	
		(*b*) Report category boundaries when continuous variables were categorized	NA	
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	NA	
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	NA	
Discussion				
Key results	18	Summarise key results with reference to study objectives	√	13-14
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	√	**16**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	√	15-16
Generalisability	21	Discuss the generalisability (external validity) of the study results	NA	
Other information				
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	NA	

* The study unit is study not participants, therefore we answered all questions pertained to patients: selection, counts, as considering study unit.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml