Observational data (Real World Data)

Subgroup report

Aldana Rosso, DK (subgroup lead, from September 2017)
Alexandra Pacurariu, EMA
Alison Cave, EMA
César Hernandez Garcia, ES
Katherine Donegan, UK
Marjon Pasmooij, NL
Martin Erik Nyeland, DK (subgroup lead, until September 2017)

See websites for contact details
Heads of Medicines Agencies www.hma.eu
European Medicines Agency www.ema.europa.eu

An agency of the
European Union

Table of content

1. Summary

1.1. *Electronic Healthcare Records data and Claims data*

Electronic healthcare records (EHRs) can be defined as an organised set of healthcare data, which can be accessed electronically. They contain a diversity of data, the most frequent being: medical records from general practitioners, specialists or hospitals, pharmacies, prescription data, and sometimes lifestyle related information.

We aimed to identify, describe and evaluate existing European EHRs for the purpose of conducting population-based observational studies to support regulatory decision making.

In the current report, a general characterisation of the data sources that were fit for purpose is provided according to a minimum set of criteria, defined by EMA staff members, which would need to be present which, includes exposure data, outcome data, type of care covered, structure and disease and drug coding, validation status and accessibility and potential for linkage.[1] Secondly, we provide a more in depth characterization of a few selected databases, which were considered appropriate for regulatory decision making, based on an initial assessment. However, there is still limited data available from some healthcare settings particularly from hospitals.

From the analysis, it is clear that the picture across Europe is patchy; some regions have extensive representation by electronic healthcare records but there are still several unrepresented regions across Europe. The number of European databases that meet minimum regulatory requirements (accessibility, validity, longitudinal data capture, both outcome and exposure recorded) and are readily accessible for use for regulatory decision making is disappointingly low resulting in a relatively low the number of patients covered in the context of the whole European population.

Moreover, mechanisms of access vary, with some data being available via commercial routes, others via academic collaborations and others only via direct collaboration with the data source holders themselves. However, there is much positivity around finding mechanisms to utilise the data for public health applications. The limitations include a substantial heterogeneity across data sources that make multi-databases studies challenging as well as complicating the comparison of results across different data sources. Moreover, there is no consistent process for understanding the validity of individual data sources for specific questions. Mechanisms for harmonising or managing heterogeneity should be sought as well as clear metrics to establish the suitability for each data source across the broad range of possible uses.

While there has been significant previous work, investigating approaches to optimise multi-database studies[2] Europe has yet to establish a sustainable network of distributed datasets to mirror that of

[1] Controlling for confounding utilising key risk factors which may be present within the dataset is critical for the use of real world data. However, the adequate controlling for confounding depends not only on the quality of information on covariates that is available but also on the methods and study design and hence is study specific and must be defined on a case by case basis. Although important, It could not therefore be considered as a general criteria for identification of databases.

[2] Coloma, P. M. et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol. Drug Saf. 20, 1–11 (2011); Trifirò, G. et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? J. Intern. Med. 275, 551–561 (2014); Bollaerts, K., De Smedt, T., Donegan, K., Titievsky, L. & Bauchau, V. Benefit-Risk Monitoring of Vaccines Using an Interactive Dashboard: A Methodological Proposal from the ADVANCE Project. Drug Saf. 41, 775–786 (2018); Eurosurveillance editorial team. ECDC in collaboration with the VAESCO consortium to develop a complementary tool for vaccine safety monitoring in Europe. Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull. 14, (2009); Mor, A. et al. Antibiotic use varies substantially among adults: a cross-national study from five European Countries in the ARITMO project. Infection 43, 453–472 (2015); Myocardial infarction and individual nonsteroidal anti-inflammatory drugs meta-analysis of observational studies – Varas-Lorenzo - 2013 - Pharmacoepidemiology and Drug Safety - Wiley Online Library. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.3437. (Accessed: 31st July 2018)

Sentinel[3], CNODES[4] and MidNet[5] limiting the ability of the European regulatory network to exploit these valuable data to complement evidence generated via other sources.

1.2. *Registry data*

Patient registries are organised systems that use observational methods to collect longitudinal, uniform data on a population defined by a particular disease, condition, or exposure. Patient registries are important sources of "big health data" evidence, particularly when combined across multiple countries, and are increasingly used throughout the pharmaceutical product life cycle. The objectives of this report were twofold: to map and characterise public registries and to discuss their application from a regulatory perspective.

Several sources have been used for the mapping and characterisation of registries: the website of the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP), the website of the EMA Registry Initiative website, and a general literature search. In this report, we describe six registries in detail: the Netherlands Cancer Registry (NCR), the Danish National Patients registry (DNPR), European Society for Blood and Marrow Transplantation (EBMT), European Cystic Fibrosis Society (ECFS), European Registry for Multiple Sclerosis (EUReMS), and the British Society for Rheumatology Biologics Registries (BSRBR). The selection of registries was based on an expert judgement to determine registries that could be used throughout the product life cycle.

The main findings of the mapping and characterisation process revealed a general lack of harmonisation with regard to data collection protocols, scientific methods and data structures across the registries. Data sharing activities between the registries are also limited although there are some excellent examples across Europe. Several recommendations are presented to improve the utilisation of registry data in the regulatory context: (i) registries should aim to standardise data fields, dictionaries and coding; (ii) governance principles and standards for transparency, accessibility and stakeholder interaction should be defined; and (iii) registries within the same disease area should exchange information to a greater extent.

The report also discusses the main areas of application of patient registries across the product life cycle. At the current time, registry data are mainly used in studies that are carried out following marketing of a product to obtain more knowledge of its safety and effectiveness (post-authorisation safety studies). In a few cases, data from registry studies has been used as an integral part of the efficacy assessment in the marketing authorisation application but the above challenges currently limit the use of patient registries data from a regulatory perspective. There is a need to facilitate access to patient registries for different stakeholders. Finally, implementation of a European standard for registry studies defining areas of applicability, data protection methods and harmonising patient consent would be beneficial to facilitate the use of registry data for regulatory decision making.

1.3. *Drug consumption data (Sales and Prescription data)*

Drug sales and prescription data databases provide information on the sales of medicines from manufacturers or wholesalers to pharmacies (community and hospital based) and retailers permitted to sell drugs, and the dispensing or sale of medicines from pharmacies to patients.

The IMI PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium[6]) project has already extensively reviewed the use, characteristics and availability of drug consumption data sources across the EU. This report reviews the mapping conducted within that

[3] https://www.fda.gov/Safety/FDAsSentinelInitiative/ucm2007250.htm
[4] https://www.cnodes.ca/
[5] http://www.pmda.go.jp/files/000223348.pdf#page=4
[6] http://www.imi-protect.eu/documents/DUinventory_2011_6_WORD97-2003.pdf

project and uses it to describe the characteristics of these types of data relevant to their use within medicines regulation. Two databases, selected to illustrate data coming from a sales database and a prescription database, are characterised in detail. A short literature survey was also conducted to identify the different uses of drug consumption data that may be of value throughout different stages of the product lifecycle.

Drug consumption data are useful tools in the regulatory process but have several limitations that may require complementary data from additional sources. There are some limitations of the data currently available, notably the limited availability of individual patient-level prescribing data particularly from hospital in-patients. Initiatives to connect in-patient hospital data should be an area of high priority. Comparison of data across different countries is also likely to be of interest and standardisation of the data across countries helps facilitate this. Finally, it is important that regulators have consistent and easy access to drug consumption data and that there is expertise available to analyse it as it can be a useful resource for routinely supporting signal assessment and for monitoring the actual or potential impacts of regulatory action.

2. Background

Big Data in Health is already being generated in a digital form and is available for use from various different sources. Vast numbers of electronic health records describing hundreds of millions of patient lives are routinely generated and collected in the context of delivering healthcare. These data have been used secondarily for many years, beyond their administrative and clinical aim, for conducting observational studies in the post-marketing stage. In Europe, such data is complemented by detailed, longitudinal patient disease registries as well as prescription/dispensing drug databases, which capture a picture of drug exposure. Today with advances in information technology, which offer the ability to access and integrate data across multiple data sources and thus generate evidence in a timely and meaningful way, it is increasingly proposed that such data can provide complementary evidence to support decision making across the product life cycle.

A number of significant challenges complicate the use of large healthcare databases. For example, the data can be both structured and unstructured and exists in many formats and terminologies. The database content is variable with time, and the quality and completeness of those sources are sometimes unknown. The need to link different databases and issues regarding accessibility to the data[7] add another layer of complexity.

In Europe, several healthcare systems co-exist and healthcare databases are not homogeneous across and within countries. Hence, there is a need to identify and describe the data in a comprehensive manner to understand its strengths and limitations and ultimately, ease its integration. An HMA/EMA Joint Task Force on Big Data was established in March 2017. The work on Phase 1 – "Mapping/characterisation of data sources" and on Phase 2- "Applicability and Usability of data sources"- by the observational data subgroup in this task force, is presented in this report.

3. Scope

The objectives of this report are:

- To identify relevant health care databases that could be of value to support medicines regulation decisions.
- To identify areas in the product life cycle in which those sources may facilitate regulatory decisions.

[7] Rijnbeek, hw 04-3 global network for her-based big data analysis, Journal of Hypertension, September 2016

- To propose a set of recommendations to better implement observational data in regulatory decision making.

3.1. *Included in the scope*

The following sources of data were reviewed in this report:

- Electronic Healthcare Records data for both primary and secondary care.
- Claims data, i.e. health data derived from insurance plan claims.
- Drug and disease registry data.
- Sales data including MAH wholesale and point-of-sale data captured under the Falsified Medicines Directive.
- Prescription/dispensing data.

Examples of data sources from different disease areas are included and specific case studies have been characterised. The applicability and usability of these data in the regulatory process is discussed. Finally, a set of recommendations for implantation of these data sources in the regulatory process is provided. A detailed description of these topics is presented in the following reports:

- Report 1: Electronic Healthcare Records data and Claims data.
- Report 2: Registry data.
- Report 3: Drug consumption data Sales and Prescription data.

3.2. *Out of scope*

The report focuses on European data sources but if relevant specific data sources outside Europe have been included for comparative purposes, but this has not been done comprehensively. Observational data from clinical trials are not included as this data source is covered by the Clinical Trial Subgroup of the Big Data Task Force. Finally, implication of data protection regulations in the applicability of the data is not discussed.

4. Reports

4.1. *Electronic Healthcare Records data and Claims data*

4.1.1. Background

The digitisation of data that is routinely generated and collected in the context of delivering healthcare has increased enormously in the last decade. Vast numbers of electronic health records are currently being collected globally describing hundreds of millions of patient lives which when coupled with advances in information technology now offer the promise to access and integrate these data and generate evidence in a timely and meaningful way. These data have been used secondarily for many years, beyond their administrative and clinical aim, for conducting observational studies in the post-marketing stage. It is increasingly proposed that such data can provide complementary evidence to support decision making across the product life cycle.

However, the use of these databases is complicated by a number of significant challenges, including the heterogeneity of the data which can exist in both structured and unstructured formats, the variable content, the quality and completeness, the need to interrogate the data across multiple disparate

sources and issues regarding accessibility to the data (Rijnbeek et al, 2016). These problems are particularly apparent across Europe driven in part by the different healthcare systems driving heterogeneous content, different legislative environments, different information technology solutions and different coding systems resulting in a lack of harmonisation across databases. Hence, there is a need to understand and describe the data in a comprehensive manner in order to ultimately, ease their integration.

This report is a deliverable from the HMA/EMA Big Data task force, Observational data subgroup and focuses on the characterisation and mapping of a subset of big data, namely electronic healthcare records (EHR). Electronic healthcare records can be classified into claims/administrative databases and medical records. Terminology in this area is often unclear with the term EHR and electronic medical records (EMR) often used interchangeable. As a result, throughout this document, we have used the definition from IMI GetReal glossary as described above (Goettsch, IMI Get Real).

Claims (administrative databases) were first established in North America in the 1980's and were the first automated databases used for population-based research. They record a person's use of the healthcare system and consist of the billing codes that physicians, pharmacies, hospitals, and other health care providers submit for reimbursement of costs to payers (Strom and Hennessy, 2012). Claims databases usually contain information from primary care, hospital and pharmacy on medical procedures, and dispensed drugs and can be linked to create a longitudinal record but frequently data protection legislation prevents linkage between different health care providers. As a result, particularly in the US, the longitudinal capture of information is often limited as patients frequently move between different healthcare providers following changes in location and employment. This is generally not the case in EU where the patients' turnover is lower and where in many countries the healthcare service is publicly funded; consequently, patient follow up is generally substantially longer and can be life-long.

Claims databases include a wide range of reimbursed expenses such as medication, hospital or GP visits, paramedical activities and lab tests. Given their primary purpose is for managing re-imbursement payments, most claims databases are frequently audited and validity checks are routinely performed for drug dispensing data, which leads to a high quality and completeness of data on drug exposure. However, the recording of medical diagnosis is less consistent, and time of diagnosis may be inaccurate as a claim for a certain diagnosis maybe made when the diagnosis has not yet been confirmed.

EMRs provide an alternative data source and represent a diverse collection of medical records from general practitioners, specialists or hospitals. EMRs contain notes and coded information collected by and for the clinicians in the office, clinic, or hospital and are mostly used for diagnosis and treatment (Ruigomez, 2002). They are longitudinal in nature and the validity of diagnosis is better than in the claims databases since this data is routinely used to inform medical care. However, as is also true for claims data, information about socio-economic factors and lifestyle choices as smoking and alcohol consumption are often lacking, as well dispensing of over the counter drugs.

These data may be obtainable through linkage to other sources such as biobanks (Hanlon, 2018) and cohort studies (Eussen, 2010) or by collection of additional data from the patients through the caregiver (van Wieren-de Wijer DB, 2008). Such approaches should be encouraged as potential approaches to build a more holistic picture of the patient.

EMRs often contain only one type of care setting (primary or secondary) although information for the other type of care may be obtained through linkage.

4.1.2. Objectives

To identify, describe and evaluate existing European databases for the purpose of conducting population-based observational studies to support regulatory decision-making. Both hospital-derived and primary care based EHRs were included as well as claims databases.

Clinical trial data sources and product or disease specific registries were considered out of scope since the data sources will be assessed by other sub-groups within the taskforce. However, if registries could be linked, to create a network of registries that effectively mimics an EHR in terms of data completeness (e.g., Nordic registries) they were included in the analysis.

4.1.3. Methods

4.1.3.1. Identification of data-sources

A team of pharmacovigilance and pharmacoepidemiology specialists identified potential longitudinal EHR and claims data sources using the ENCePP repository but also data sources identified in web-based search engines for healthcare databases (e.g. BRIDGE TO DATA), textbooks on clinical pharmacoepidemiology and from EU funded research projects such as PROTECT, ADVANCE, GRIP, EU-ADR, FP7 Drug Safety research programmes and other non-EU regulatory initiatives (e.g. FDA Sentinel). Data sources listed under the EMA framework contract for post-authorisation effectiveness and pharmacoepidemiology research were also included. Initially publicly available information about each data source was retrieved and reviewed to extract key information and data source owners or governance entities were later contacted to provide detailed feedback on the data collection process, data characteristics, exposure, population coverage, linkage, data access conditions and validation studies.

Following the initial identification and high-level characterisation, the data sources were further screened to determine whether they contained data, which could potentially support regulatory decision-making. For the included data sources, information from publicly available sources was supplemented with a survey sent to all database owners. Information requested within the survey included database characteristics, population coverage and the number of active patients included in the database; inquiry about linkage with other data sources and how it may be possible, conditions for access and level of access possible and information on any validation studies performed using the data source (the survey is provided at Appendix I). The survey sent was sent to 63 institutions and the response rate for the survey was 81%.

Additionally, a series of teleconferences were organised to clarify some of the information provided within the responses or when the data owners requested further information around the underlying reasons for the request.

As a preliminary step, data sources were excluded from further analysis if any of the below exclusion, criteria were met:

- External collaboration was not possible; data holders were asked whether it was possible for an external organisation (government/academic research organisation) to access the database.
- There was no longitudinal data capture.
- Only exposure or outcome data was captured[8]. Prescription only databases[9] were excluded based on the consideration they cannot be used for full etiological studies; however, we acknowledge

[8] An exception from this rule was when the database can be routinely linked to other data sources and mimic a full EHR, for example the national network of registries in Nordic countries, which are traditionally linked and used together
[9] Prescription databases are covered by a separate subgroup

their usefulness for drug utilisation studies which are often undertaken to inform regulatory decisions and understand current clinical practice.

- The data source was a drug or disease specific registry[1].
- The dataset was no longer active and historical data was not accessible.

For the remaining databases, further analysis was undertaken.

It is important to emphasise that the focus of this analysis was on identifying data sources to support regulatory decision making across a broad number of use cases e.g. high quality of recording of exposure to medicines and therefore exclusion on the basis of the above criteria does not necessarily reflect the quality or completeness of the data and the excluded databases might be relevant for other types of studies. Moreover, different data sources may serve different regulatory purposes, but the scoring was binary and was not intended to capture this complexity.

4.1.3.2. Validation

Database owners were asked to report the validation studies for their database of which they were aware. 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 valid information from another database capturing the same information for a different purpose.

4.1.3.3. Coding

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. The datasets were coded in each of the following domains.

4.1.3.3.1. The extent of data capture

Demographic data:

- Number of patients.
- Number of active patients.
- Paediatric patients.
- Patient age or year of birth.
- Patient gender.
- Pregnancy data or possibility of pregnancy identification through parent child linkage.
- Other socio demographic factors as deprivation scores were not investigated.

4.1.3.3.2. Exposure to medicines[10]

- Posology of the medicine.
- Duration of treatment (if recorded per se or can be calculated from other existing information).

[10] No distinction was made between prescription vs dispensing information.

- Route of administration.
- Inclusion of vaccinations (administration or prescribing information).
- Clinical diagnosis recorded as indication or a proxy thereof.

4.1.3.3.3. Health outcomes and additional healthcare related data characteristics

- Clinical diagnosis.
- Screening tests.
- Laboratory test results.

4.1.3.3.4. Validity

- The presence and number of existing validation studies[11].

The extent to which a database was validated and considered fit for the conduct of observational studies was investigated indirectly, through the number of validation studies published in the scientific literature. Database owners were asked to report validation studies of which they were aware, 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 valid information from another database capturing the same information for a different purpose.

4.1.3.4. Accessibility and linkage

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.

- Accessibility of data and level of access provided.
- External linkage possible to other datasets.

4.1.3.5. Data transformation

Data holders were asked whether the database had been transformed into a general common data model (CDM) or was in process of transformation. A CDM transformation provides a common representation of the data across multiple databases thus enabling the standardisation of administrative and clinical information and facilitating a combined analysis across several databases. The CDM may be systematically applied to all structured data within a database ('full' CDM e.g. OMOP), to a subset of data ('partial' CDM e.g. Sentinel) or to a subset of data needed for a specific study ('study-specific' CDM). One dataset could be transformed into more than one CDM.

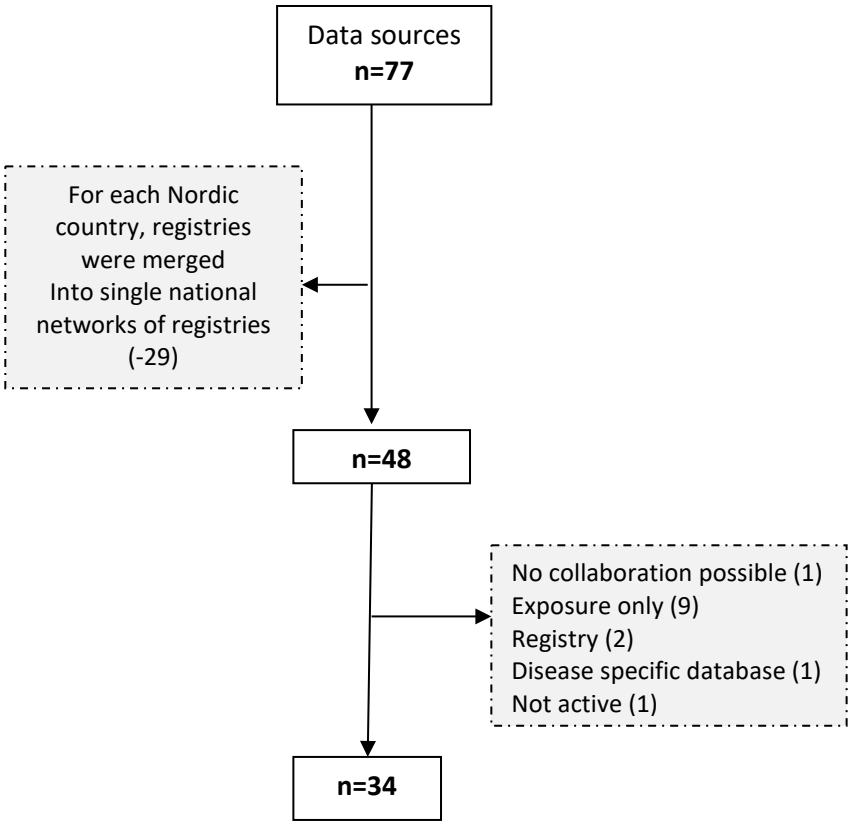
The underlying assumption is that a robust and validated CDM will serve a regulatory purpose, for example, it would accelerate an analysis. However, this does not eliminate the role of studies in individual databases where a more in-depth analysis can be done as well as extra validation studies.

[11] By validation study we define any study that attempted to check the accuracy of a recorded variable (either diagnosis or prescription) by comparison with a reference standard.

4.1.4. Results

The initial search generated a list of 77 potential data sources (see Appendix 2A for listing). The Nordic registries were clustered in one single entry per country for easier retrieval (from 33 to 4). Out of the remaining 44 data sources, 13 were excluded on the basis of at least one of the main exclusion criteria: only exposure data (8), being categorised as registries (2), no possibility of collaboration (1), the database was no longer active (1), and restricted to a disease specific database (1) (see Figure 1). The final number of data sources selected was 34 (see Appendix 2B for full list).

Figure 1. Data sources selection flowchart.



Thirty-four sources were retained, and their basic characteristics are described further below.

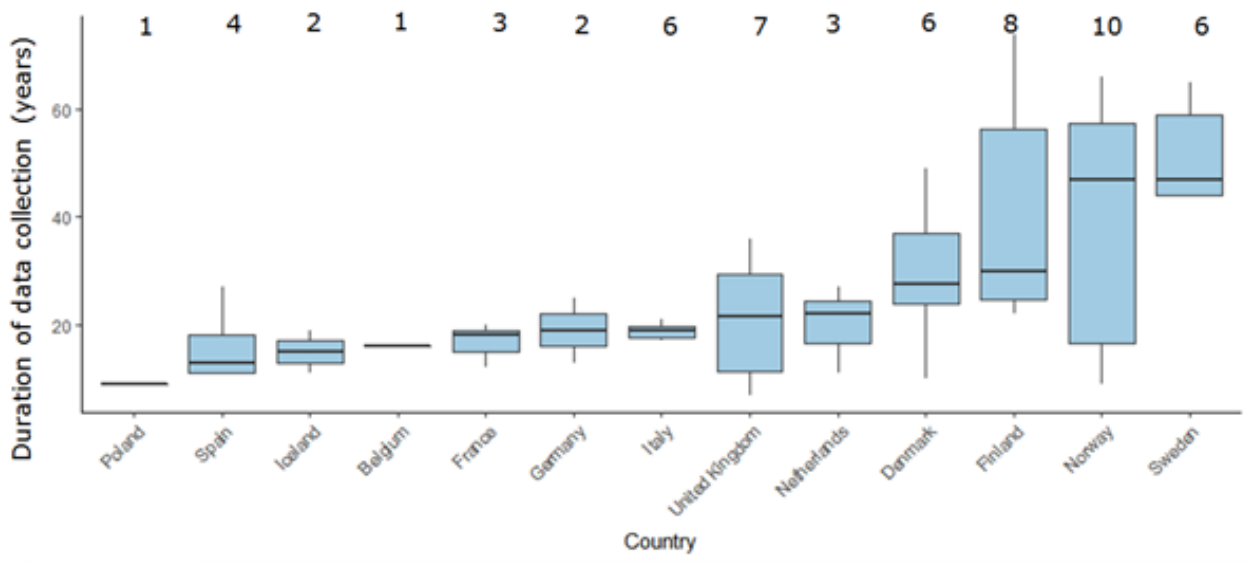
4.1.4.1. Volume Size of the data source

The median number of total patients across the datasets is 5 million patients (range 0.07-15 million). However, the number of active patients is generally lower, 13% databases having more than 10 million patients and 52% of databases having more than 5 million patients.

Basic demographic variables as age and gender of patients are recorded in all (100%) of data sources. Paediatric data is included in 28 data sources (90%). At the current time no detailed analysis of the extent and quality of the paediatric data present in the datasets has been performed which is acknowledged to be very variable.

The period covered by the majority of databases spans from 1990 to present day with the oldest data source being established in 1964 (the National Finnish Hospital Discharge Register). The median calendar time covered by a database is 18.5 years (range 7-53 years) (see Figure 2).

Figure 2. Number of data sources across Europe and duration of data collection.



Each boxplot represents the distribution of duration for all databases in that country. The numbers above the box represent the number of individual databases within a country.

In terms of geographical coverage, 17% of databases come from Norway, 14% from Finland and 10% from Denmark and Italy (see Figure 2).

4.1.4.2. Structure

We classified the included data sources, based on structure, purpose and type of data contained in the following categories: electronic medical records, claims databases or healthcare record linkage system (when multiple databases are used together through linkage). The most frequent type is electronic medical records (38.7%) and record linkage system (32.3%). Claims/administrative databases comprise only 19% of the EU databases.

In order to better understand the data that may be contained in EMRs there is a need to understand the clinical care pathways in each country, for example how and which type of care is delivered in each setting, the presence or not of a gatekeeper system in primary care, delivery of paediatric care and how preventive measures are delivered and recorded. This landscaping work is currently underway at the EMA for each member state.

A range of care settings are covered by the included data sources, with most of them including mixed care settings (primary and secondary care). However often data from secondary care is of variable coverage and quality.

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

Table 1. Cross tabulation of data sources and type of care covered

4.1.4.3. Exposure related information

All included databases contain information about exposure to a drug (either prescribed or dispensed) as a prerequisite for inclusion in the inventory. 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 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 are not consistently captured in the databases while vaccinations were captured in 13 databases (38%). Data on hospital in-patient administered drugs were rarely captured (8.8%). Outpatient used drugs and drugs administered in hospitals are captured depending on the type of care covered (see Section 4.1.4).

4.1.4.4. Outcome related information

All included databases have information about medical events (outcomes) as a prerequisite for inclusion in our inventory. The completeness and quality of information is however variable, as well as the way of recording the information: 23 (74%) of the databases record medical events as diagnosis or symptoms. The remaining databases provide the possibility of linkage to hospital registries or other sources, which contain outcome data.

4.1.4.5. Other recorded information

Referrals for laboratory investigations were captured in in 20 (59%) and referrals for imaging or other diagnostic procedures were captured in 17 (50%). databases. Within the scope of this analysis we did not distinguish between those databases that simply record referrals for procedures and tests and those, which also record the results of the investigations; however, the number of those who record the outcome of the procedures is expected to be much lower. Recording of lifestyle factors was not included in this analysis.

4.1.4.6. Veracity

Data provenance

Health care data is collected at regional or national level from either hospital or general practice patient chart review, pharmacy records or administrative records. The data is generated in the course of delivery of normal clinical care or for payment purposes not for the purposes of research and therefore will be dependent upon a number of variables which include the motivation of the healthcare professional entering the data, the requirements of the organisation, the audit processes in place, software tools to prevent missing data fields or erroneous data and the type of consent provided by the patient.

Quality

Quality was not systematically assessed during the creation of the inventory.

Completeness

The completeness of the information was variables across data sources and was evaluated at a high level with the aid of a scoring algorithm. The scoring algorithm was designed to enable filtering of the databases on a number of key characteristics (see appendix 3 for links to the spreadsheet). Score 1 is computed based on data source characteristics namely collaboration, longitudinal data, recording of exposure and recording of clinical events. Score 2 refers to data source elements such as size of the data source, access to and analysis of data, linkage potential, presence of hospital data and paediatric

data, the patient characteristics and disease characteristics included, validation studies and the potential transformation of the data to a common data model. Score 1 ranges from 1 to 6 while score 2 ranges from 1 to 21 points. In general, datasets with the highest scores were more comprehensive and likely to be of interest for regulatory decision-making. As a caveat, it can be legitimately argued that the value of a database depends on the research question to be answered and there is no 'absolute' value applicable for all studies. Therefore, feasibility of a specific database should always be considered on a case-by-case basis (Hall GC, 2012).

4.1.4.7. Representativeness

The databases population are usually highly representative for the source population covered, be it regional or national. The coverage range is from 3% (Pedianet) to 100% (Nordic registries, HSE Ireland, eDRIS-ISD, SAIL, Vektis and ARS) with an average coverage of 60%.

The representativeness of the data at a European and worldwide level depends on many factors and was not evaluated within the scope of the current analysis.

4.1.4.8. Analytical tools

For the purpose of multi-database studies, is possible to analyse the data either using a common protocol or by extracting the data in a common data format. For a review of the data management and analysis technique, please see Bazelier et al (2015) and Chapter 4.6 Research networks from ENCePP Guide on Methodological Standards in Pharmacoepidemiology.

4.1.4.9. Accessibility and potential for linkage

Only one database was excluded because no third-party access was allowed. From the remainder, 32% 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 through use of PINs. Other forms of linkage were sometimes used. For example, in order to avoid the use of PINs 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 (21%).

4.1.4.10. Validation

No published validation study was reported for 17 (50%) databases, 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). Validation studies are usually outcome specific and thus do not validate the entire data source for every type of studies. The validation concerned either specific health outcomes or prescription information. Some database owners have reported as validation studies, validation of prediction algorithms for various health outcomes as chronic kidney disease, ischaemic stroke and various types of cancers based on estimating the absolute risk of a particular outcome in primary care patients with and without symptoms. It is debatable if these are truly validation studies according to our definition. Validity should be judged and investigated on a case-by-case basis before initiating a study.

4.1.4.11. Variability Data heterogeneity

There is significant heterogeneity between the databases as reflected by the overall score; nevertheless, there are some core elements that are present in almost every database (see Appendix 3).

4.1.4.12. Data standards

Data standards exist almost always at database level. No data standards at an EU level exist for observational data as is the case for RCT data. The broad scope and complexity of health information make implementation of standards challenging.

Various coding systems and ontologies are employed by different databases. The most widely adopted coding systems are ICD and SNOMED, according to a WHO survey on eHealth.(5)

4.1.4.13. Data processing

The amount of data processing that may occur within a database to facilitate data sharing and the conduct of pharmacoepidemiological studies was not included within this assessment e.g. the imposition of mandatory fields within an EHR model or of specific data ranges for laboratory results or age. However, we included a field to highlight whether data holders had transformed the data or were considering transforming the data into a common data model in order to facilitate multi-database studies as this is relevant for the support of timely and representative studies across Europe.

4.1.4.14. Transformation of the databases to a common data model (CDM)

A CDM transformation provides a common representation and architecture of the data across multiple databases thus enabling the standardisation of administrative and clinical information facilitating a combined analysis and the use of common analytical tools across several databases. The CDM may be systematically applied to all structured data within a database ('full' CDM e.g. OMOP), to a subset of data ('partial' CDM e.g. Sentinel) or to a subset of data needed for a specific study ('study-specific' CDM'). One dataset could be transformed into more than one CDM. Any transformation to a CDM is likely to involve a certain degree of information loss, especially when a full CDM is used. The outstanding question is whether the information loss affects the interpretation of the study results or whether it was unnecessary detail. The advantage of using a CDM is that transformed databases are more accessible for research across a network and could therefore increase the speed and power of multi-sites studies. This may be particularly important from a regulatory perspective if there is an urgent safety issue or a rare exposure or outcome. Hence, we recorded either the actual transformation of a data source into a CDM, whether performed by the data holder themselves or by a third party, or the intention of the data holders to transform their database into a CDM.

Across the final 34 selected databases, four databases had transformed their data and five others were in the process of transforming the database.

Table 2: Databases converted to CDM and type of model used.

Database name	Country	CDM model type	Status
QuintilesIMS Disease Analyser	France	OMOP[12] CDMv5	Complete
QuintilesIMS Disease Analyser	Germany	OMOP CDMv5	Complete

[12] http://omop.org/node/608

Spanish Information System for the Development of Research in Primary Care	Spain	ADVANCE[13]/ OMOP CDM	In progress
Pedianet	Italy	OMOP CDMv4	In Progress
Agenzia Regionale di SanitàTuscany database	Italy	OMOP CDMv5	In progress
Clinical Practice Research Datalink	United Kingdom	OMOP CDMv5	Complete
Integrated Primary Care Information Database	Netherlands	OMOP CDMv5	In progress
The Health Improvement Network	United Kingdom	OMOP CDM	Complete
Information System of Parc de Salut del Mar	Spain	OMOP CDM	In progress

OMOP= Observational Medical Outcomes Partnership; CDM=Common Data model; ADVANCE= Accelerated development of vaccine benefit-risk collaboration in Europe

4.1.4.15. Velocity Speed of change

No information on the speed of change of the datasets was undertaken in terms of the incorporation of new data elements into the databases such as genomic data, imaging data or patient reported outcomes. It is clearly envisaged that integration of such data with EHRs will be undertaken in the future and there are specific initiatives where this is already occurring e.g. integration of UK Biobank phenotyping data with healthcare records. Such linkages bring opportunities but will create multiple challenges around not only managing the volume of the data and achieving standardisation across multiple databases but also around enabling machine learning solutions on data held in multiple sites. Recognising these challenges, BBMRI-ERIC[14] has created a European research infrastructure for biobanking offering support with quality management (including international biobanking standards, auditing, and training), legal, ethical and societal issues, open source online tools and a directory of biobanks.

4.1.4.16. Rate of accumulation

No information was collected on the rate of accumulation of data within the databases. However, it is clear that the rate of data accumulation is an increasing challenge especially if unstructured data held within many of the data sources is considered. In order to exploit such data innovative centralised approaches will be required to integrate, store and mine data to generate novel insights.

4.1.4.17. Value

All observational datasets such as EHRs contain many uncertainties including but not limited to multiple terminologies, confounders (both known and unknown) and have variable architectures, content, completeness and quality. Thus, many scientific decisions need to be taken when undertaking a study and it is essential that a researcher has an in depth understanding of the strengths and limitations of the datasets to avoid potential erroneous and misleading results. As advocated by Wang

[13] https://www.imi.europa.eu/content/advance
[14] http://www.bbmri-eric.eu/

et al (2017) at a minimum, transparency around study protocols and operational definitions used to create the analytical dataset would substantially help to increase confidence in the validity and robustness of the evidence generated via studies in healthcare databases.

4.1.5. Regulatory Challenges

Over the last decade, the scientific landscape has changed dramatically creating regulatory challenges, which are demanding new approaches to the generation of complementary evidence across the product life cycle. For example, new insights generated by genomics are challenging the way we think about disease enabling greater disease and thus patient stratification; 'omic technologies are generating new biomarkers offering the opportunity to diagnose disease at an earlier stage enabling early intervention but also opening up new methods to accurately track treatment response; the availability of the electronic health records of millions of patients and improvements in data analytics offer new opportunities to understand the usage, effectiveness and safety of medicines in clinical practice and the digital phenotyping revolution on our mobile phones may in the future allow the incorporation of patient centred outcomes into our decision making processes.

Hence there are increasing opportunities to utilise the data but currently, greater applicability is limited by multiple challenges which range from a fundamental need to establish appropriate access to the data, to the need for new analytical methods to enable the integration and analysis of heterogeneous datasets and the generation of meaningful conclusions. Moreover, we need to understand the limitations in the data to know where it can offer the most value for which we need appropriate and robust validation of the datasets and confidence in the methodology used to generate the evidence especially when studies are conducted across multiple datasets. For observational studies the ENCePP Code of Conduct provides a set of rules and principles for pharmacoepidemiology and pharmacovigilance studies to promote transparency and scientific independence throughout the research process[15].The Code is aimed at permitting a high level of public scrutiny which ultimately will increase the confidence of the general public, researchers and regulators in the integrity and value of pharmacoepidemiology and pharmacovigilance research. To this end, the Code promotes best practice standards for the interaction of investigators and study funders in critical areas such as planning, conduct and reporting of studies.

As a core transparency measure, and whether or not studies fully comply with the Code, the protocols of all pharmacoepidemiology and pharmacovigilance studies should be registered in the EU PAS Register ideally before they start, and study findings should be published irrespective of whether the results are positive or negative. Above all compliance with data protection legislation and robust and transparent mechanisms to protect patient confidentiality are critical to securing patient trust.

4.1.5.1. Data Accessibility

The accessibility of European datasets for research/third parties was a prerequisite for retention in our inventory; 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%).

Additional value arising from data linkage, current approaches for data linkage and limitations

In terms of linkage, 68% of the databases can be linked through a unique personal identification number (PIN) to other databases which contain additional healthcare related information such as cause of death registries, hospital data, prescription databases or cancer registries or be used to enrich the EMR with lifestyle and other information by linking to biobanks and well phenotyped cohorts. This

[15] http://www.encepp.eu/

offers opportunities to significantly extend the applicability and usability of the datasets but does create challenges in ensuring data privacy in terms of the risk of patient identification. The Nordic registries are a good example of the extensive use of linkage among different registries by usage of PIN, forming national record linkage systems. Other forms of linkage are sometimes utilised. For example, in order to avoid the use of PIN to preserve anonymity the PHARMO network uses a probabilistic linkage based on patient birth date, gender and GP code. The linkage of a parent with their child ('parent-child linkage'), which is useful for studies investigating pregnancy exposures and effect on offspring, is available in only 7 data sources (22.5%).

4.1.6. Key case studies

Five data sources are presented in more detail as case studies.

4.1.6.1. The Health Improvement Network

Background

The Health Improvement Network (THIN) is a primary care database, which was set up in 2002. In December 2016 over 730 practices had contributed data, with a total of 15.6 million patients of which just over 3 million active (currently registered) patients can be prospectively followed. The database covers approximately 6% of the UK population and is considered representative for the source population.

Basic demographics such as age, sex, birth date are recorded, although complete birth dates or other identifiers are not available in order to preserve confidentiality. All medical conditions and symptoms recorded during the GP consultation appear in the THIN database and are coded with READ codes.

GP prescriptions are also recorded and are of good quality since they are sent electronically to the pharmacy for the dispensation of drugs. Drugs are recorded using Anatomical Therapeutic Chemical Classification (ATC) terminology. THIN contains information on lifestyle and preventative healthcare, including variables such as height and weight measurements, blood pressure (BP), smoking and alcohol status, immunisations and lab test results (so called Additional Health Data) however the degree of completeness is variable across contributing practices and lower than for main characteristics.

THIN is able to provide anonymous postcode linked area based socioeconomic, ethnicity and environmental indices (PVIs) at a patient level. Secondary care information is now available through Hospital Episode Statistics (HES) database, which is linked to THIN Data. HES data holds details of all hospital admissions, outpatients, accident and emergency attendances, maternity care and critical care at NHS hospitals in England.

4.1.6.2. The UK healthcare system

The NHS was set up in 1948 and is the world's largest publicly funded health service. Currently the NHS has a workforce of over 1.6 million people, including doctors, dentists, nurses, midwives, managers, ambulance staff and therapists. The NHS is free at the point of delivery for anyone who is resident in the UK (currently over 64 million people). It covers everything from routine treatments for coughs and colds to heart surgery, accident, emergency treatment, and end-of-life care. The NHS is divided into two sections: primary and secondary care. Primary care is the first point of contact for most people and is delivered by a wide range of independent contractors, including GPs, dentists, pharmacists and optometrists. Secondary care is known as acute healthcare and can be either elective care or emergency care. Elective care means planned specialist medical care or surgery, usually following referral from a primary or community health professional such as a GP.

4.1.6.2.1. Accessibility

The THIN database is available for research, subject to a fee. In the UK, all research involving data collected from National Health Service (NHS) patients must be approved by a Research Ethics Committee (REC). REC has approved the THIN data collection scheme as a whole and has permitted the establishment of a scientific committee to review protocols for scientific merit and feasibility. The protocol for studies for publication conducted using THIN data must be approved by Scientific Review Committees (SRCs) which has also to approve the data collection scheme. If a protocol is approved, access to patient level data is granted.

4.1.6.2.2. Published articles

Over 500 peer-reviewed studies have been published using THIN data. An overview is available at the following link https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/publications.

4.1.6.3. Clinical Practice Research Datalink

4.1.6.3.1. Background

CPRD is managed by the Department of Health, United Kingdom and is a governmental, not-for-profit research service. It is a primary care database, which was set up in 1987 and contains medical records of 22 million patients, representative of the general UK population. Over 10 million active (currently registered) patients can be prospectively followed.

The CPRD contains data regarding demographics, symptoms and diagnosis, tests, immunisations, interventions, referrals to secondary care. In terms of drugs exposure, it contains prescriptions issued in primary care, with usual information on formulation, strength and dosing instructions available. CPRD datasets contain structured and coded data. The key coding schemes and dictionaries used in the NHS are ICD-10, READ, OPCS4, SNOMED CT and the British National Formulary (BNF).

Lifestyle information (e.g. smoking and alcohol status) is also present in various degrees.

CPRD can be linked with datasets from secondary care, disease-specific cohorts and mortality records to enhance the scope for research. CPRD has full access to Hospital Episode Statistics (HES) data, which can be made available as separate modules of hospitalised care, outpatient visits (visiting a consultant), maternity care and augmented/critical care.

4.1.6.3.2. Accessibility

Data is available for research and a fee-basis. The CPRD has an internal research team which offers support for protocol development, gaining approvals for research, data extraction and analysis and medical writing for reports and publications.

4.1.6.3.3. Published articles

Over 1,700 articles published in peer-reviewed journals have used data from the CPRD.[16]

4.1.6.4. The Information System for the Development of Research in Primary Care

4.1.6.4.1. Background

SIDIAP includes data collected since 2005by health professionals during routine visits in primary care centres throughout Catalonia, including clinical diagnoses, laboratory tests, prescribed and dispensed

[16] https://www.cprd.com/bibliography/

drugs, immunisations, hospital referrals, mortality, demographic and lifestyle information. It contains anonymised data for nearly six million people (5.588.922 in December 2015) which represents approximately 80% of the Catalan population. This is considered to be representative of the actual conditions in clinical practice and Garcia-Gil Mdel et al (2011) demonstrated that the SIDIAP population is highly representative of the entire Catalan region in terms of geographic, age, and sex distributions. Hospital discharge information is also available for those patients treated in a Catalan Health Institute hospital (approximately 30% of the SIDIAP population).

SIDIAP allows linkage with other databases in Catalonia at an individual level through a mechanism that guarantees the confidentiality of the clinical data. Linkages include those with the Cancer Registry of the Hospital del Mar, the Arthroplasties registry, and the Girona Dementia registry.

Several validation studies have been performed in SIDIAP. These studies have used different strategies to validate the registered information in SIDIAP, such as the comparison of the SIDIAP data with a gold standard e.g. a cancer registry to validate cancer diagnosis or comparison of incidences of health problems in SIDIAP with the incidences reported by other sources of information, such as cohort studies.

4.1.6.4.2. Accessibility

Investigators of research groups accredited by the 'SIDIAP Jordi Gol research foundation' or investigators of primary care of the Catalan Health Institute can have direct access to the SIDIAP data. Investigators of research groups from other public research institutions or regulatory authorities can also apply to obtain data from the SIDIAP via a signed agreement. To determine individual access conditions to the data, SIDIAP takes into account various aspects, including data safety.

SIDIAP does not provide data to for profit organisations. However, the SIDIAP can conduct research projects and deliver a report of the results to such organisations at the end of the investigation. Under this scenario, a SIDIAP research group would define the design of the study together with the external entity and after approval by the Scientific and Ethical Committees of the SIDIAP, the SIDIAP research team would carry out the investigation and deliver the different reports as stated in the signed agreement between the SIDIAP Jordi Gol research foundation and the external organization.

Data can be used for research purposes only and there is no transfer of raw data to the study sponsor. SIDIAP is responsible for the study design and analysis of data with the delivery of a final study report. Research projects of the research groups accredited by the SIDIAP are considered of high priority by the SIDIAP and are charged a lower fee for data access compared to other public research institutions or for profit organisations.

4.1.6.4.3. Published articles

Over 50 peer-reviewed studies have been published using SIDIAP[17].

4.1.6.5. Finnish Record linkage system

4.1.6.5.1. Background

The Finish record linkage system contains a set of registries, which are linked through a personal identification number:

- National Hospital Discharge Register/ Care Register for Health Care and Social Welfare.

[17] http://www.sidiap.org/index.php/dissemination/articles

- Causes of Death Register Finland.
- Finnish Linked National Health Registers.
- Finnish Prescription Register.
- Medical Birth Register.
- National Hospital Discharge Register.
- Register for Congenital Malformations.
- Register for Induced Abortions.
- Register of Primary Health Care Visits.

Since all administrative databases include the personal identity code as a mandatory variable, linkages between different registers are feasible. A good example of such a linkage exercise is the Finnish Linked National Health Registers at THL.

The registries cover the entire population of Finland, approximately 5 million patients. As the names of individual registries suggest, they contained information about diagnosis in primary care or secondary care, hospital admissions, laboratory data. The National Hospital discharge register is the oldest database of this kind in Europe, being established in 1967.

Medical Birth Register contains data on diagnoses during pregnancy and birth as well as for newborn and can be linked with medications dispensed during pregnancy[18], which makes it a very valuable data source for etiological studies examining exposure and during pregnancy and outcomes in newborns.

The register for Congenital Malformations contains information on major congenital anomalies detected during pregnancy or before the age of one year among live births, stillbirths with a gestational age of 22+0 weeks or a birth weight of 500 grams or more, or termination of pregnancies.

All these registers are population based and cover all patients. The reporting is based on completed care excluding the registers on cancers, congenital malformations and visual impairments. There is no information on active patients, but the data of any year can be linked to subsequent years and to deaths (Cause-of-Death Register at Statistics Finland, permission for data linkage is required).

The legislation on secondary use of health and social welfare data in Finland is being updated. This proposal would allow more liberal use of the register data, including for example commercial and non-commercial development work.

4.1.6.5.2. Accessibility

Data from a single register or multiple registers can be used for scientific or historic research. The register keeping organisation can grant a permission to use the confidential register data for scientific purposes if the predefined conditions are reached: that is if the study plan is scientifically sound and justifiable, the study questions can be answered with the exiting data, the study is led by a scientific merited person, and all data protection rules and regulations are fulfilled. Only universities and other scientific research centres can receive such permission. Permission to access data includes a fee in addition to a fee for the time needed for the preparation for the data will be charged. If data from different register keeping organisations are to be used, the application process must be done for each organisation separately. More information on the procedure is available at the THL website[19].

[18] https://www.thl.fi/fi/web/thlfi-en/research-and-expertwork/projects-and-programmes/drugs-and-pregnancy
[19] http://www.sitra.fi/en/well-being/well-being-data

The SITRA fund (operating directly under the Finnish Parliament) is funding several projects, which are developing better solutions for register-based research and analyses, e.g. by streamlining the application process for permissions and single access point for register data and for distance access solutions for register-based research and data analysis.

4.1.6.5.3. Published articles

No overview available.

4.1.6.6. IMS Disease Analyser

4.1.6.6.1. Background

This is dataset captures healthcare data from primary care in Germany. Data are entered by primary care practices and specialists from a random sample of practices, which are weighted by region, doctor specialty, and size of community and age of doctor in order to achieve a representative sample of the population. Records are available from 1992 and are updated monthly. Approximately 3.2% of all doctors in Germany (3,300 physicians from 2,700 practices), over the past 3 years have been recruited capturing about 15 million patients in DA Germany. However, in Germany, patients are not registered with a single doctor, which means activity status needs to be evaluated during the analysis otherwise misclassification might occur. As a result, the dataset is not suitable for the study of long-term effects.

Basic information is included on age, gender, height, weight, BMI and smoking. Medical events are recorded as ICD-10 codes. Lab test results are available, but procedures are not routinely captured. Further information is collected on date of visit, symptoms, risk factors, adiposity, comorbidities, referrals and hospitalisation. Test results and diagnostic tests are included in the dataset for practices who are connected to a laboratory that capture data electronically. Prescribing information contains molecule, brand & generic name, manufacturer, form, date, pack size, strength, dose, cost, and insurance type. Drugs are coded via WHO and EphMRA ATC codes and supplemented by text fields. No remedies and aids are recorded.

Validity and representativeness of the database was investigated and shown to be adequate (Becher, 2009).

4.1.6.6.2. Accessibility

The dataset is available for research on a fee basis. The conditions and data license fees for this and other IMS data sources vary widely according to the country of origin, the sample size of the study population, etc.

4.1.6.6.3. Published articles

Published articles are available here[20].

4.1.7. Conclusions

There is a wide range of health care databases available for epidemiologic research in Europe, some of which are very well-established and with a long tradition in electronic recording of medical data. The most common data sources available are record linkage systems with a mix of primary and secondary care coverage. Many of these databases have been used for regulatory decision making in the past,

[20] http://www.quintiles.com/experts/publications-by-quintiles-authors

mainly for safety related questions post authorisation and it is envisaged that this data may have broader application. However the number of European databases that meet minimum regulatory requirements (accessibility, validity, longitudinal data capture, both outcome and exposure recorded) and are readily accessible for use for regulatory decision making is disappointingly low resulting in a relatively low number of patients covered in the context of the whole European population. Moreover, the datasets are geographically restricted with the most represented regions being Northern, Central and Western Europe with very few databases from eastern European countries.

Over 60% of databases have internal validation processes and various validation studies in the peer reviewed literature, which generates some reassurance in the quality of the data source. However, a comprehensive validation should normally be study-specific, since it depends on the study objective. Thus, in itself the number of validation studies performed cannot be used as an indicator of the overall validity of the database but may nevertheless provide some confidence in the data source and inform researchers on the feasibility to perform study-specific validation in individual database.

Not all of databases are accessible for research, although accessibility increases if an organisation can comply with ethical and scientific requirements. The access for a third party such as EMA, for research purposes, is provided at patient level data in 24% of cases while the remaining datasets have more restrictive access policies.

One of the biggest challenges in using these data sources for research arises in the context of multi database research due to the high heterogeneity in coding and structure. While not all questions demand this approach, it is clear that there are multiple factors which influence the benefit-risk of a medicine and some of these may be country specific. Hence generating data across multiple countries in a timely manner is a key requirement. A possible solution to address these issues may be via the transformation of data into a common data model which standardises structure and in some models terminology; in Europe work to explore the feasibility of this approach is ongoing but still in its early stages. However, in order to use data transformed into a CDM, a robust and systematic validation of the transformed data against the sources data would be required which would need to be maintained over time.

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. It is appreciated that it is difficult to define a priori which databases may be suitable to answer a research question of regulatory interest, as requirements will be study specific and hence be variable. Furthermore, the above review has a number of limitations. Firstly, some data sources may have been missed during the identification process. However, in an attempt to be as complete as possible several rounds of database identification were incorporated and the inventory was reviewed by experts, including members of the ENCePP Working Group “Data sources” and database owners. Where possible data from publicly available sources was complemented or verified with database owners. A number of difficulties were encountered when trying to map all the existing EHDs in Europe, which highlights again the need for more comprehensive and accessible repositories with EHDs. Secondly, prescription only databases were excluded since they cannot be used for etiological studies although it is acknowledged 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 and this was only evaluated indirectly through the number of validation studies reported by the database owners.

4.1.8. Recommendations

- There is still limited health care data available from secondary care or from specialist care settings across Europe.[21] Given many new medicines are prescribed in this setting, including innovative medicines, mechanisms are needed for the collection, standardisation and harmonisation of secondary care/specialist care data. In the future, such data sources may be utilised in the context of pragmatic trials, recognised in the US by the establishment of PCORnet.[22]
- Creation of a maintained, central inventory of data sources, detailing general characteristics and characteristering the strengths and limitations would help identify suitable data sources across a broad range of regulatory questions with often very variable requirements.
- There is a clear need for the development of data sources in European member states, which currently either have no data sources or are poorly represented.
- Mechanisms for combining data across European data sources should be implemented to increase timely access to observational data. There are several possible approaches including transformation of data into a CDM, which may be full, partial, or study specific. The most appropriate approach for European data sets remains to be decided and was the topic of an EMA workshop in 2017.[23]
- The integration of new data sources within EHRs should be supported. Improved linkage across records is required to deliver a holistic picture of the patient health status. Standard terminologies and methodologies are needed to enable the incorporation of data from novel data sources e.g., m-health, and patient reported outcome measures in a consistent and validated manner.
- Recording of information about exposure is variable especially for route of administration. Implementation of ISO IDMP standards within EHDs would enable the unique identification of the medicinal product including brand, batch number, dose, and route of administration. This would be particularly important for biologicals given the recent evidence, which suggests this batch number is poorly recorded within observational data (Klein et al, 2016).
- Referrals and results of laboratory values are commonly missing from primary care records. However, laboratory tests provide valuable quantitative information and mechanisms to more consistently record the timing and outcomes of laboratory tests would add significant value.
- Implementation of a mandatory recording of indications of use, outcome measures and cause of death would increase utility for regulatory focussed questions.
- Sustainable mechanisms should be sought to promote collaboration and facilitate consistent and timely regulatory access to data sources.
- Complete transparency with regards to validation studies and conduct of more validation studies should be encouraged. The development of robust validation measures and an increased transparency of validated outcomes would improve consistency and replicability of studies across different databases.

[21] Disease specific data from secondary care may be available via patient registries (for a review see following report)
[22] PCORnet consists of 20 Patient-Powered Research Networks (PPRNs), 13 Clinical Data Research Networks (CDRNs), 2 Health Plan Research Networks (HPRNs) and 1 Coordinating Centre. For further information please visit http://www.pcornet.org/
[23] http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/events/2017/10/event_detail_001524.jsp&mid=WC0b01ac058004d5c3

4.1.9. Regulatory applicability across the product life cycle

The term electronic healthcare data sources (EHDs) encompasses both electronic healthcare records data and claims/administrative data and are frequently used in variety of settings to study the use of drugs and associated health outcomes. In the previous sections, we have outlined a number of their advantages including: larger size, which allows the study of infrequent events, increased representativeness of routine clinical care, their availability at a relatively low cost and the significantly decreased time to complete a study because the data are already collected and available.

However, concerns about EHD based studies centre on data validity, lack of details about lifestyle factors and socio-economic factors, and a limited ability to control confounding at data collection stage (Schneeweiss and Avorn, 2005). Challenges also exist in terms of accessibility to data in a timely fashion particularly in certain care settings.

As for all observational studies, acceptability of the evidence arising from the use of EHDs in the regulatory setting depends upon a number of factors including:

- the potential and feasibility of capturing other data,
- the delay likely to be imposed by the request to generate additional data by other means (e.g. to perform a clinical trial or to validate diagnoses in medical records),
- the unmet medical needs,
- how well progression of disease is currently understood by the investigators,
- what is known about the benefit-risk of the product to understand whether the effect size for the intended study would likely be discernible in EHD,
- the likely characteristics of the patient population in the EHDs and the ease of identifying a consistent study population (are there clear reproducible, precise biomarkers of disease likely to be recorded in the EHD),
- the ability to accurately record exposure and,
- the presence of a hard end point likely to be recorded consistently and accurately in EHD.

Moreover, acceptability will also be influenced by the existing methods to account for potential bias and what is already known about the quality of the data source. All of these factors will be influenced by the regulatory setting in which the evidence is intended to be used.

4.1.10. Use of EHDs in the pre-authorisation phase

The use of EHDs in the pre-authorisation phase is currently very limited, partly due to the fact that a prerequisite for the generation of RWD is for the drug to be marketed; hence there is very limited historical use and regulatory experience, around the use of RWD to support effectiveness decisions. However, it is conceivable and there are rare examples where RWD albeit derived from patient registries, has been used as a source of historical control data for a newly marketing authorisation application. However, RWD has been used to contextualise other evidence, for example to provide an understanding of the natural history of the disease, current clinical care and unmet needs and enable calculation of incidence/prevalence measures for the purpose of designation of orphan status. A few examples are presented below.

4.1.10.1. Approval of an orphan medicine - the use of historical controls

It is perhaps in the area of orphan medicines where the case for evidence generation outside of RCTs may be most compelling. One recent example is Zalmoxis, an orphan gene-therapy product, which recently received a conditional approval as an adjunctive, or add-on, treatment for adult patients receiving a haplo-identical haematopoietic stem cell transplant (HSCT) with high-risk haematological malignancies to aid immune reconstitution and reduce the risk of graft-versus-host disease.[24] An ongoing RCT will deliver results in several years but it was considered that waiting for these results represented an unacceptable delay in an area of unmet medical need. Hence the conditional approval was based on a small single arm study (57 patients), the results of which were compared with the control arm of the ongoing Phase III trial combined with controls selected on the same criteria derived from the European Bone Marrow Transplant Registry. Such an approach resulted in uncertainties around the impact of the differences in baseline characteristics between the historical, concurrent controls and treatment arms and hence resulted in a conditional approval pending the completion of the ongoing Phase II trial.

While this is an important example of the application of RWD, it is doubtful that EHDs could provide the level and quality of data needed for such a comparison, particularly in the setting of a rare disease.

4.1.10.2. Recruitment of patients for RCTs through EHDs

The EHR4CR project,[25] completed in 2016, involved 35 academic and 11 hospital sites across Europe ("Electronic Health Records for Clinical Research - (EHR4CR)," n.d., p. 4). The project has developed a platform that can utilise de-identified data from hospital EHR systems, to assist clinical trials feasibility assessment and patient recruitment. The platform can connect securely to the hospital EHR systems and clinical data warehouses across Europe, to enable a trial sponsor to predict the number of eligible patients for a candidate clinical trial protocol, to assess its feasibility and to locate the most relevant hospital sites. In addition, recently, machine-learning techniques have been applied to routinely collect patient data from electronic health databases to develop algorithms, which could be used to identify currently undiagnosed patients for specific diseases, useful for RCT recruitment but also a potential key need to explore efficacy of medicines intended to prevent the onset of disease. For example, Doyle et al (2017) used US prescription and open-source medical claims between 2010 and 2016 to capture information on demographics, treatments, procedures and symptomatology, comorbidities/misdiagnoses and specialist visits for patients diagnosed with hepatitis C virus (HCV) and no HCV. Machine learning approaches were then employed to develop algorithms, which identified medical differentiators for HCV, which occurred in the years prior to the diagnosis of HCV.

- Pragmatic clinical trials

A pragmatic clinical trial is 'a study comparing several health interventions among a randomised, diverse population representing clinical practice, and measuring a broad range of health outcomes' (IMI Get Real Glossary). They are focused on evaluating treatments in patient populations and settings more representative of routine clinical practice. There is increasing interest in this approach as such trials address a criticism of RCTs in terms of the unknown generalizability of results to clinical practice and yet can still incorporate randomisation within the trial design. A recent well known example of a pragmatic clinical trial is the Salford lung trial, which was a Phase III pragmatic RCT where patients were enrolled through primary care practices using minimal exclusion criteria and without extensive or non-routine diagnostic testing (Collier et al., 2017). The safety outcomes were then captured through patients' electronic health records and revised by the specialist safety team. There were significant

[24] http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Summary_for_the_public/human/002801/WC500212516.pdf
[25] http://www.ehr4cr.eu/

challenges in establishing the linked network to deliver this trial and it is inefficient to create such extensive networks for single trials. As such, global initiatives have been launched aimed at developing networks of EHDs to enable the conduct of efficient pragmatic trials on a routine basis. A key example is PCORnet which is a large, highly representative, national patient centred clinical research network[26] in the US encompassing 79 distinct health systems with representation in every state, with the vision of enabling high quality, efficient large scale clinical research both observational and interventional. No equivalent network exists in Europe.

4.1.10.3. Orphan medicines

Understanding the diseases or condition for which the drug is indicated can be particularly challenging for orphan drugs, for which the prevalence of the disease or condition must not be greater than more than 5 in every 10,000 individuals. For an orphan designation the company must demonstrate prevalence in the European Community and moreover prevalence figures should be from more than two countries (Hall and Carlson, 2014). To support their authorisation, the demographic parameters of disease prevalence and the quantification of trends in incidence and prevalence over time is needed but due to the rarity of the condition it may be challenging or impossible to use classical sources such as literature or expert opinion and real world data sources such as EHDs may be helpful. However, the rarity of the conditions may mean that a single data source would not provide sufficient patients to accurately estimate prevalence and thus again networks of data sources such as the US Sentinel system may be required to deliver sufficient cases. Specific disease registries will also be extremely helpful in providing a clear understanding of disease history, any disease stratification, potential geographical differences as well as diagnostic criteria and procedures across the disease course. However providing an estimate of prevalence may not be possible from a registry due to an inability to define a denominator for the dataset.

One example of an orphan drug which was approved based on limited clinical evidence is Ninlaro® (ixazomib) with an indication for relapsing multiple myeloma ("Ninlaro Assessment Report," n.d.). The product received a conditional approval in a specific group of patients (multiple myeloma who have received at least one prior therapy); based on the interim results of the pivotal Phase III randomised controlled trial even though 'the efficacy evidence was not as comprehensive as normally required'. The conditional approval was because the product was aimed at a life-threatening condition, the interim results showed a positive risk-benefit and it was likely the applicant will be able to provide comprehensive data. Final results from the ongoing RCT will be due in December 2019 and will be complemented by additional RCTs and by an observational clinical study (NSMM-5001) which will further describe treatment patterns and patient outcomes in 1000 patients in order to complement and contextualise the current evidence on efficacy ("Ninlaro Assessment Report," n.d.).

These few examples demonstrate there is still a limited application of RWD pre-authorisation but there is huge interest among all stakeholders to understand how this data may be better utilised. This interest is partly driven by the challenges presented by a changing scientific landscape but also by IT innovations, which are offering new possibilities for the creation of distributed data sources. However, because of the multiple uncertainties in the application of RWD in the regulatory setting, it has, currently only proved acceptable when the unmet medical need is clear and there are no alternative feasible mechanisms to capture the data. The challenge for us all is how to create sufficient trust to allow this data, which may bring potentially novel insights, to be used more routinely. However perhaps because Europe has arguable some of the greatest challenges in harmonising data sources across multiple languages and terminologies, it is lagging behind North America in the creation of

[26] PCORnet consists of 20 Patient-Powered Research Networks (PPRNs), 13 Clinical Data Research Networks (CDRNs), 2 Health Plan Research Networks (HPRNs) and 1 Coordinating Centre. For further information please visit http://www.pcornet.org/

established data networks to better enable the secondary use of health data in a routine and consistent manner. Other approaches to enable re-use of healthcare data should be considered such as the standardisation of terminologies across care settings to allow the incorporation of clinical trial outcome measures into RWD. Such an approach would facilitate a proactive assessment of benefit-risk following authorisation supporting the refinement of the label over time.

4.1.11. Use of EHDs in post-authorisation phase

EHDs use in post-authorisation phase is better established and evidence from such data has supported regulatory decision making for many years. There are multiple potential applications which include population description and treatment patterns exploration, causality assessment for safety signals, genome-wide association studies or assessing the effectiveness of risk minimisation measures. (Schneeweiss and Avorn, 2005).

4.1.11.1. Extension of indication for an old drug based on bibliographical references

There may be potential for the use of RWD to support an extension of an indication for an already marketed product and there is significant interest in understanding where and when such data may be acceptable from the regulatory context. For most new medicines, the regulatory process usually demands an RCT to provide an unbiased estimate of efficacy. However, there have been circumstances where medicines have been approved by regulatory authorities without randomised evidence. One example is provided by the approval of a new indication for an old product 6-thioguanine (6-TG), based on evidence from RWD (van Asseldonk et al., 2011) (Fraser et al., 2002).

6-TG has been authorised for the treatment of leukaemias since 1975 but it was observed that the product was being used off-label at a much lower dose for second line treatment of inflammatory bowel disease in patients not responding to or intolerant to other treatments. Significant and prolonged off label use suggests considerable unmet need in this patient population. Given its long history of authorisation there was sufficient non-clinical evidence available and no further evidence was required. Thus, the MAH applied for a new indication based on bibliographic evidence and a single bioequivalence study comparing the new product Thiosix 20 mg versus the established product Lanvis 40 mg. scientific advice was sought by the MAH to understand, in the absence of an RCT, the dossier requirements to achieve authorisation for this indication. Despite the fact there were no studies with a randomised controlled design using the proposed or other formulations of 6-TG, no new clinical studies were performed. The supportive evidence was derived from 11 uncontrolled studies (prospective uncontrolled studies in a clinical research setting and retrospective database studies with data captured from screening of medical records from various hospitals) to support the efficacy of the proposed dose of 20mg/day in the target population (Meijer et al., 2016). The submitted studies overall included 307 patients, two thirds of whom were intolerant to AZA/6-MP but were heterogeneous in design; five studies had a median duration of 6-9 months while two studies had a median follow-up up to 22 months. There were several other uncertainties associated with the studies; they were primarily conducted in Western Europe (the Netherlands, Germany, France, UK), and therefore the vast majority of the subjects were likely to be Caucasian. Additionally, response and/or remission rates varied between 35% and 89% and different measures were used for efficacy outcomes. It was concluded that a robust estimation of efficacy versus placebo or established treatment was lacking and that additional data was needed (CBG-MEB, 2015). Consequently, while Thiosix was granted a conditional approval for short-term therapy, given the outstanding concerns regarding the efficacy and safety of the proposed posology for long-term maintenance, the creation of a registry was required to deliver an understanding of the long-term effectiveness. Matched historical controls will be provided for a comparator group.

It is clear that there was significant and prolonged off label use of 6-TG which suggests considerable unmet need in this patient population, and it is likely that this was an important driver in considering an authorisation of a product based on bibliographic evidence despite the risk of publication and other biases. In addition, the presence of an authorisation enables the requirement for further long-term studies to explore the benefit-risk of the product and is likely to stimulate the reporting of adverse drug reactions. It is also worth highlighting that in circumstances where bibliographical evidence from EHDs and mitigating circumstances e.g. unmet need were not sufficient for a marketing authorisation, findings such as these would support the conductance of randomised trials to demonstrate efficacy.

A second example where RWD has been used to support effectiveness claims is eculizumab, whose indication was extended on the basis of real world evidence from the global registry established by the company at the time the product was first marketed as part of its post marketing obligations.[27] Acceptability of RWE in this case was driven by the rarity of the disease and the efficacy of eculizumab that made a non-treated arm unethical. Ultimately this data allowed a comparison of outcome in patients with no transfusion history treated or not with eculizumab and enabled an extension of the indication to patients with haemolysis with clinical symptoms indicative of high disease activity regardless of transfusion history. In this case deriving sufficient robust data was challenging even when the source was a disease registry, established by the MAH in order to meet post-marketing obligations. It is highly unlikely that sufficiently robust data could have been captured from EHDs alone.

A third example is ivacaftor, a medicine designed to target a specific mutation in the cystic fibrosis transmembrane regulator gene, whose indication has been extended multiple times on the basis of non-randomised evidence predominantly due to the fact that the mechanism of action and target of the medicine was completely understood.[28] It may be expected that such extensions will occur for other genomically targeted medicines where similar mutations to the original target mutation are identified following the original authorisation. The challenge for the regulatory community is to determine the level of evidence that will be required under these circumstances to allow an extension of the indication to encompass further mutations.

4.1.11.2. Signal detection and management

There is an increasing interest in exploring the use of EHRs for signal detection especially under situations when health outcomes are hard to identify from spontaneous reports (e.g., myocardial infarction) or outcomes, which may be temporally dissociated from the initial exposure. In these cases, EHDs might provide a better data source and as such, there are a number of initiatives looking to develop new approaches to better utilise EHDs for signal detection. For example, IMI PROTECT performed an evaluation of the usefulness of EHDs for signal detection and explored the opportunities and challenges for prospective signal detection, compared options for exploratory and confirmatory analysis and evaluated the performance of longitudinal data for quantitative signal detection compared with individual case study reports. They concluded that while longitudinal data should be further explored as a complement to signal detection via spontaneous ADRS, none of the positive drug-event pairs could be detected in EHDs at an early stage. In addition, prospective signal detection in EHDs should include clinical, pharmacological and epidemiological review of potential signals and if possible, explored using statistical graphical methods to remove false positives (Cederholm et al, 2014). Signal detection in EHDs should also account for the limitations in the underlying data, in particular its size and scope, to ensure appropriate interpretation. Lastly future research should explore the relative merits of performing signal detection for groups of products and medical events in the same class as

[27] http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Summary_for_the_public/human/000791/WC500054210.pdf
[28] http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Summary_for_the_public/human/002494/WC500130744.pdf

commonly done versus individual products and events. It is common for epidemiological studies to be performed for all drugs in a class together and/or for a number of related medical events together to improve power. However, a detailed review revealed substantial and important differences among different products in the same class or among different medical events in the same category (Wisniewski et al, 2013) emphasising the need for individual review.

A further example of the application of novel techniques within EHDs is provided by a software application package Treescan.[29] Treescan is a data mining method which implements the tree-based scan statistic within administrative claims data, and simultaneously looks for excess risk in any of a large number of individual cells as well as in groups of closely related cells, adjusting for the multiple testing inherent in the large number of overlapping groups evaluated. It has a number of potential applications:

- to simultaneously evaluate hundreds or thousands of potential adverse events and groups of adverse events, to determine if any one of them occur with higher probability among patients exposed to a particular pharmaceutical drug, device or vaccine, adjusting for the multiple tests inherent in the many adverse events evaluated,
- to simultaneously evaluate if a particular disease outcome such as liver failure occurs with increased risk among people exposed to any of hundreds of pharmaceutical drugs, or groups of related drugs, adjusting for the multiple testing inherent in the many drugs evaluated,
- to evaluate whether certain occupations, or group of related occupations, are at higher risk of particular diseases.

Equally, EHDs could add significant value by supporting signal validation through the provision of information about drug utilisation, additional risk factors which may influence the incidence of a particular ADR or understanding the causal relationship between a potential drug-event pair. Advantages of EHRs for signal validation are:

- providing exposure measures to put the relative risk in context.
- testing of the potential association through etiological studies.
- investigating confounders.
- providing more clinical context of the target population (what are the additional risk factors, main co-administered drugs, etc.).

A further consideration is the availability of appropriate codes in EHDs for the detection of ADRs. For example, no ICD 10 code exists for osteonecrosis of the jaw and thus Ehrenstein et al (2015) evaluated the positive predictive power of an algorithm based on a number of ICD9 and 10 codes and concluded the predefined algorithm was not adequate for monitoring of ONJ for pharmacovigilance studies. Additional case finding approaches coupled with adjudication, would be necessary to increase confidence in detection highlighting the need for development of coding to allow for code-based detection of ADRs. There is also a need to maintain up to date mappings between MedDRA and coding terminologies used in observational data.

4.1.11.3. Post Authorisation Studies

Secondary use of routinely collected data from electronic healthcare records and claims databases in post authorisation studies is popular for several reasons: it is usually faster and cheaper than primary data collection, has the potential to access large patient populations, provides the opportunity to access data from a wider geographical area and it has increased external generalizability compared to

[29] (https://www.treescan.org/).

primary data collection (Schneeweiss and Avorn, 2005). A review of 189 PASSs 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 cases. Among the 19 (33%) PASS which used secondary data collection, 58% leveraged electronic health records (EHRs) (Engel et al., 2017). A second review of studies registered in the EU PAS Register, found that 117 studies (37%), used an existing claims or electronic medical records database. (Carroll et al., 2017)

Drug utilisation studies were more likely than other type of studies to use secondary data. There is a difference in application across disease areas most likely related to the reliability of recording relevant outcomes in the EHDs; as such EHDS are more commonly used in type 2 diabetes mellitus, COPD and cardiovascular disease database studies (Carroll et al., 2017). However, availability of data on exposure in EHDs is often incomplete, especially start date, duration of exposure, dose and adherence to treatment and prescribing is rarely linked to a specific indication. More consistent recording of indications of use, outcome measures and cause of death would significantly increase utility.

4.1.11.4. Effectiveness of risk minimisation measures

Use of real-world data is essential in order to assess the impact of risk minimisation measures and cannot be replaced with data collected in a controlled environment. Real-world data is essential for evaluation of both drug utilisation and health outcomes, the latter being directly linked with the public health impact.

A review of pharmaceutical industry-sponsored studies evaluating the effectiveness of risk minimisation measures submitted to the EMA for cardiovascular, endocrinology or metabolic drugs authorised between 1995 and 2015 found that 42% of studies evaluating routine and additional risk minimisation measures used EHDs. (Mazzaglia et al., 2017) Another review of 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 former being the most utilised type of data sources for such studies. (Goedecke et al., 2017) However, there are currently challenges in accessing data representing the whole of Europe hampering the determination of whether risk minimisation measures are equally effective in different member states. As such, our recent analysis could only identify 34 EHDs across 13 member states, which appear relevant for regulatory decision-making.

4.1.12. Regulatory acceptability of the data

As regulatory decisions based on EHDs may have a major impact on public health, the quality of the information contained in the databases and the validity and reproducibility of the derived results are critical. EHDs have been used as part of the evidence package for many years to support pharmacovigilance decisions where the opportunities to capture other data are often limited. However, there remains a significant concern that evidence derived from real world data cannot meet the evidentiary standards required to support regulatory decisions on efficacy and effectiveness. These concerns stem partly from unknowns about the data quality and partly due to non-random allocation of treatments and subsequent unknowns around the extent of confounding by indication.

The above analysis has focussed on the data itself to guide whether a data source may be useful for regulatory decision-making. However a number of recent publications have illustrated that both the choice of database and the methodological design can have a profound impact on the derived evidence which have added to the concern around the ability of observational studies to deliver robust evidence for regulatory decision making but in particular for the determination of effectiveness where other opportunities to capture data may be available. For example Madigan et al (2013) demonstrated the potential impact of database choice on observational study results by systematically studying

heterogeneity across 10 databases and 53 drug outcome pairs and 2 widely used epidemiological study designs (cohort and self-controlled case series) (Madigan et al., 2013b). The authors demonstrated that despite holding study design constant, 20%–40% of observational database studies can swing from statistically significant in one direction to statistically significant in the opposite direction depending on the choice of database. This exceeded the proportion of pairs that were consistent across databases in both direction and statistical significance. While the approach has limitations in that the same methodological approach may not be appropriate for all drug-outcome pairs (Gruber et al, 2016) it nevertheless illustrates the importance of study design. As such in a further analysis, Madigan et al (2013a) demonstrated that clinical studies using observational databases were sensitive to both study design and to specific analytical choices within the design; applying alternative study designs to an investigation of a supposedly negative association between bisphosphonates and four health outcomes not only demonstrated that different design yielded discrepant results but moreover the influence was different for different outcome measures. Klungel et al (2016) similarly examined the consistency of findings from different drug-outcome pairs across multiple designs and databases and different European countries; in contrast to results of Madigan et al (2013a) these authors demonstrated that while there was some variation in the magnitude of the effect size, it was consistent in direction across multiple designs, databases and methods to control for confounding and none of the differences were statistically significant.

The question therefore arises as to how potential drug effect signals arising from observational data may be verified. In an attempt to develop a 'reference standard' against which to test the ability of a database to return a true finding, OMOP researchers created a reference database of 'positive' and 'negative' drug event combinations that should be expected to return the appropriate response from a signal detection test (Ryan et al, 2013). A key regulatory need is to be able to quickly determine whether a signal is positive or negative. A later publication (Schuemie et al., 2014, 2018) demonstrated that results returned positive associations across a number of drug-outcome pairs presumed to represent negative controls; as such, across 30 negative controls between 57% and 73% associations revealed risk estimates that were either significantly harmful or protective which casts doubt on any statistical significant result generated by observational data. Recent publications suggest that calibration of P values against negative controls may be necessary to improve the reproducibility of observational studies but the whole community remains to be convinced (Gruber and Tchetgen, 2016). However, a limitation of this approach is that identifying a perfect negative control is challenging; as such, Hauben et al (2016) reassessed the negative controls identified by OMOP researchers and found problems with 17% (40 of 233) of the classifications.

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 (Hall et al., 2012). 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.

In addition to known biases and confounders associated with the methodology, the potential limitations of EHRs/claims databases that might impact in regulatory field are:

- Validity: information about the validity of the data within the data sources is scarce and usually limited to specific outcomes and/or treatments.
- Completeness of records - either in terms of proportion of individuals that are captured in the data source, missing variables (as lifestyle factors or laboratory test values) or the comprehensiveness of a record. Since the collection of data is not under the control of the researcher, there is no possibility to address missing data other than at design stage. In the future better data linkage between records e.g. between hospital and primary care or between health records and pregnancy,

education, social or tertiary care records may provide corroborative evidence in the event of inconsistencies. However, such linked systems are currently only available in Scandinavia and due to the richness of the information are not easily accessible.

- Misclassification of diagnosis, exposure and outcomes - misclassification of exposure or outcome will be present to an unknown amount in EHRs and moreover is likely to vary depending on the outcome and the investigated data source, the sensitivity of the disease and should be taken into consideration when interpreting the results. Different epidemiological study designs will also be variably affected by misclassification (Funk and Landi, 2014).
- The size of data sources- although by virtue of size, most EHDs should have more potential statistical power than clinical trials and other primary data sources, there are instances in which the size of a specific database might still be too small e.g. rare diseases, exposures or outcomes, for the evidence to be conclusive. In these cases, multi-databases studies are recommended.
- Data accessibility – while a percentage of EHDs are becoming commercialised, many are still not easily accessible and the procedure for access can be onerous and lengthy especially for multi-database studies. This limits utility for regulatory decision making where deadlines can be tight particularly for addressing urgent safety signals.

4.1.13. Solutions for improving regulatory acceptability

Collaboration among all stakeholders, regulators, industry, healthcare professionals, academia and database owners is required to increase acceptability of these data sources for regulatory decision-making. Given the number of issues, areas of focus with the potentially greatest impact need to be developed but should include:

- Understanding and documenting the validity of EHDs - regulators should encourage the conduct of validation studies for specific databases, approaches and outcomes. For example, more consistent validation of case finding algorithms should be requested to assess the extent of misclassification and estimate its impact on the study results. Development of more robust validation measurement accepted and routinely applied by the community would aid replicability across studies. It is anticipated that such measures will ultimately drive an increase in quality of the data sources, but regulators need to articulate clearly their needs. The EU Scientific Advice procedure offers an opportunity for interaction between data holders and the EU regulatory network, which will ultimately deliver greater understanding of the limitations and enable pragmatic agreement to be reached. The procedure provides a process by which the data holder may receive scientific advice and potentially a CHMP qualification opinion or advice around for example specific methodologies to harmonise data sources or around the suitability of specific data sources for regulatory decision making (European Cystic Fibrosis Society, 2017). Some scientific journals also launched an appeal to researchers to fill this gap. ("Pharmacoepidemiology and Drug Safety - Wiley Online Library," n.d.) (Ehrenstein et al, 2016).
- While single database studies are appropriate for some questions, certain scenarios such as orphan diseases, rare exposures, rare safety events or special populations require access to more cases. As such, the performance of multi-databases studies and the creation of distributed networks of databases or research networks should be encouraged e.g. ENCePP. ("ENCePP Resources Database," n.d.). Approaches such as common data models should be explored in depth as mechanisms to enable timely access to data across multiple databases. However, it is considered that no one approach will be suitable for regulatory needs and a hybrid approach will always be required.

- Increasing the representativeness of EHDs for the European setting by encouraging the development of EHDs in MS currently underrepresented.
- Improving accessibility –database owners should be encouraged to increase both the ease and speed of accessibility, at least in case of regulatory requests. In this context, providing the ability to access aggregated results in a manner which protects patient privacy is likely to have significant impact.
- Increase timely access to data – regulatory decisions particularly in the context of safety decisions may need urgent access to data from across Europe to inform regulatory decision-making. Currently accessing data across multiple datasets via a common protocol method can take many months to agree a protocol, access the data, and provide results to inform decision-making. Common data models (CDMs) to harmonise data structure, terminology and dictionaries enable common analytical methods to interrogate and extract results across datasets transformed into the CDM providing results in a timely manner. However, there are several CDMs currently utilised and it is key for the regulator to identify which CDM best balances timeliness with sufficient flexibility to address regulatory questions. Moreover, any CDM should incorporate robust and transparent validation processes and there is a need to develop robust business models, which would support the transformation and maintenance of data transformation of European datasets.[30]
- Raising awareness about database selection guidelines that would help investigators to select from the start databases that are acceptable for regulators. We need to ensure that such guidelines encompass factors relevant for the European setting where the use of multiple coding systems is common. The availability of a central, maintained characterisation of European data sources based on a consistent set of parameters would enable researchers to select the most appropriate database for the question and also deliver transparency in the selection.
- Implementing a routine transparency of reporting which should include a clear justification for the choice of database and study design. Changes to the protocol should be carefully documented and justified. Registration of post-authorization studies in the EU PAS Register®, which allows uploading the study protocol, study report and publication, is recommended as it provides public access to evaluations carried out on specific drugs and specific safety or effectiveness concerns, and visibility on investigators, data availability, methods and funding sources.
- Robust data governance mechanisms to ensure data privacy obligations are met.
- Effective communication of the value of EHDs for public health activities to both healthcare professionals and patients would help to promote a data sharing culture and additionally may improve the quality of the imputed data. This would also be facilitated by feedback of clinical data in an accessible manner, which would support their clinical decision-making.
- Provision of regulatory guidelines to support the use of real-world evidence in regulatory decision-making.

4.1.14. Conclusion and recommendations

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 in supporting regulatory decision-making (Engel et al, 2017). These data sources are already a critical piece of the pharmacovigilance jigsaw picture as illustrated in the sections above and already the data provides context for many additional decisions. The changing scientific landscape is however creating further regulatory challenges where the use of EHDs could help to reduce uncertainties at authorisation but additionally provide

[30] See report from the EMA Common data model workshop in December 2017.

mechanisms for proactive assessment of long-term safety. In addition, there is increasing interest in utilising these data sources beyond pharmacovigilance and more specifically to understand the effectiveness of medicines in the real world given the unknown external validity of the majority of randomised control trials. However, while there is huge value locked away within the data sources, there is also a variability of quality and content which must be managed to build evidence of sufficient robustness, reproducibility and replicability for regulatory decision making. The series of measures described above would help to deliver a better understanding of the results from observational studies enabling the development of clear criteria, which need to be met for the studies to reach acceptable standards for regulatory decision-making. Above all in order to meet the need for timely and robust data for urgent safety issues, the development of a sustainable network of European databases is required.

4.1.15. References

Becher H, Kostev K, Schröder-Bernhardi D. Validity and representativeness of the "Disease Analyzer" patient database for use in pharmacoepidemiological and pharmacoeconomic studies. Int J Clin Pharmacol Ther. 2009 Oct;47(10):617-26.

Carroll, R., Ramagopalan, S.V., Cid-Ruzafa, J., Lambrelli, D., McDonald, L., 2017. An analysis of characteristics of post-authorisation studies registered on the ENCePP EU PAS Register. F1000Research 6, 1447. https://doi.org/10.12688/f1000research.12198.2

CBG-MEB, 2015. Public Assessment Report Scientific discussion Thiosix [WWW Document]. URL https://db.cbg-meb.nl/Pars/h114680.pdf (accessed 2.26.18).

Cederholm, S., Hill, G., Asiimwe,A., Bate,A., Bhayat,F., Persson Brobert, G., Bergvall,T., Ansell,D., Star,K., Noren,G.N. Structured assessment for prospective identification of safety signals in electronic medical records: Evaluation in the Health Improvement Netwoek. Drug Saf, 2015,38:87-100.

Collier, S., Harvey, C., Brewster, J., Bakerly, N.D., Elkhenini, H.F., Stanciu, R., Williams, C., Brereton, J., New, J.P., McCrae, J., McCorkindale, S., Leather, D., 2017. Monitoring safety in a phase III real-world effectiveness trial: use of novel methodology in the Salford Lung Study. Pharmacoepidemiol. Drug Saf. 26, 344–352. https://doi.org/10.1002/pds.4118.

Ehrenstein V, Gammelager H, Schiødt M, Nørholt SE, Neumann-Jensen B, Folkmar TB, Pedersen L, Svaerke C, Sørensen HT, Ma H, Acquavella J. Evaluation of an ICD-10 algorithm to detect osteonecrosis of the jaw among cancer patients in the Danish National Registry of Patients. Pharmacoepidemiol Drug Saf. 2015 Jul;24(7):693-700.

Ehrenstein, V., Petersen, I., Smeeth, L., Jick, S.S., Benchimol, E.I., Ludvigsson, J.F., Dorensen, H.T. Helping everyone do better; a call for validation studies of routinely recorded health data. Clin Epidemiol, 2016: 12:49-51.

Electronic Health Records for Clinical Research - (EHR4CR) [WWW Document], n.d. URL http://www.ehr4cr.eu/ (accessed 3.13.18).

ENCePP Resources Database, http://www.encepp.eu/encepp/search.htm (accessed on April 2018)

Engel, P., Almas, M.F., De Bruin, M.L., Starzyk, K., Blackburn, S., Dreyer, N.A., 2017. Lessons learned on the design and the conduct of Post-Authorization Safety Studies: review of 3 years of PRAC oversight. Br. J. Clin. Pharmacol. 83, 884–893. https://doi.org/10.1111/bcp.13165

European Cystic Fibrosis Society, 2017. Qualification Opinion for the European Cystic Fibrosis Patient Registry [WWW Document]. URL

http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2018/02/WC500243542.pdf (accessed 2.27.18).

Eussen SR, de Jong N, Rompelberg CJ, Garssen J, Verschuren WM, Klungel OH. Effects of the use of phytosterol/-stanol-enriched margarines on adherence to statin therapy. Pharmacoepidemiol Drug Saf. 2010 Dec;19(12):1225-32. doi: 10.1002/pds.2042. Epub 2010 Oct 4.

Fraser, A.G., Orchard, T.R., Jewell, D.P., 2002. The efficacy of azathioprine for the treatment of inflammatory bowel disease: a 30-year review. Gut 50, 485–489. https://doi.org/10.1136/gut.50.4.485

Funk, M.J., Landi, S.N., 2014. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. Curr. Epidemiol. Rep. 1, 175–185. https://doi.org/10.1007/s40471-014-0027-z

Garcia-Gil Mdel M. et al., Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAP). Inform Prim Care 19, 135-145 (2011).

Goedecke, T., Morales, D.R., Pacurariu, A., Kurz, X., 2017. Measuring the impact of medicines regulatory interventions - Systematic review and methodological considerations. Br. J. Clin. Pharmacol. https://doi.org/10.1111/bcp.13469

Goettsch, W. Glossary of Definitions of Common Terms, IMI Get Real.

Gruber, S., Chakravarty, A., Heckbert S.R., Levenson, M., Martin, D., Nelson, J.C., Psaty, B.M., Pinheiro, S., Reich, C.G., Toh, S., Walker, A.M. 2016 Design and analysis choices for safety surveillance evaluations need to be tuned to the specifics of the hypothesized drug-outcome association. Pharmacoepidemiol Drug Saf.; 25(9):973-81.

Gruber, S., Tchetgen, E.T. Limitations of empirical calibration of p-values using observational data. Statist Med., 2016, 35:3869-3882.

Hall, A.K., Carlson, M.R., 2014. The current status of orphan drug development in Europe and the US. Intractable Rare Dis. Res. 3, 1–7. https://doi.org/10.5582/irdr.3.1

Hall, G.C., Sauer, B., Bourke, A., Brown, J.S., Reynolds, M.W., LoCasale, R., Casale, R.L., 2012. Guidelines for good database selection and use in pharmacoepidemiology research. Pharmacoepidemiol. Drug Saf. 21, 1–10. https://doi.org/10.1002/pds.2229

Hanlon, P., Nicholl, B.I., Jani, B.D., McQueenie, R., Lee, D., Gallacher, K.I., Mair, F.S. 2018. Examining patterns of multimorbidity, polypharmacy and risk of adverse drug reactions in chronic obstructive

pulmonary disease: a cross-sectional UK Biobank study. BMJ Open 2018;8:e018404. doi:10.1136/bmjopen-2017-018404.

Hauben, M., Aronson, J., Ferner, R. 2016 Evidence of misclassification of drug-event associations classified as gold-standard 'negative controls' by the Observational Medical Outcomes Partnership (OMOP). Drug Saf. 2016. Drug Saf; 39(5):421-32.

Klein, K., Scholl, J.H., Vermeer, N.S., Broekmans, A.W., Van Puijenbroek, E.P., De Bruin, M.L., Stolk, P. Traceability of Biologics in The Netherlands: An Analysis of Information-Recording Systems in Clinical Practice and Spontaneous ADR Reports. Drug Saf. 2016 Feb;39(2):185-92.

Klungel, O.H., Kurz, X., de Groot, M.C.H., Schlienger, R.G., Tcherny-Lessenot, S., Grimaldi, L., Ibáñez, L., Groenwold, R.H.H., Reynolds, R.F., 2016. Multi-centre, multi-database studies with common

protocols: lessons learnt from the IMI PROTECT project. Pharmacoepidemiol. Drug Saf. 25 Suppl 1, 156–165. https://doi.org/10.1002/pds.3968

Madigan, D., Ryan, P.B., Schuemie, M., 2013a. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. Ther. Adv. Drug Saf. 4, 53–62. https://doi.org/10.1177/2042098613477445

Madigan, D., Ryan, P.B., Schuemie, M., Stang, P.E., Overhage, J.M., Hartzema, A.G., Suchard, M.A., DuMouchel, W., Berlin, J.A., 2013b. Evaluating the impact of database heterogeneity on observational study results. Am. J. Epidemiol. 178, 645–651. https://doi.org/10.1093/aje/kwt010

Mazzaglia, G., Straus, S.M.J., Arlett, P., da Silva, D., Janssen, H., Raine, J., Alteri, E., 2017. 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. https://doi.org/10.1007/s40264-017-0604-4

Meijer, B., Mulder, C.J., Peters, G.J., van Bodegraven, A.A., de Boer, N.K., 2016. Efficacy of thioguanine treatment in inflammatory bowel disease: A systematic review. World J. Gastroenterol. 22, 9012–9021. https://doi.org/10.3748/wjg.v22.i40.9012

Ninlaro Assessment Report [WWW Document], n.d. URL http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/003844/WC500217623.pdf (accessed 2.7.18).

Pharmacoepidemiology and Drug Safety - Wiley Online Library [WWW Document], n.d. URL http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1099-1557 (accessed 2.7.18).

P. Rijnbeek, HW 04-3 GLOBAL NETWORK FOR HER-BASED BIG DATA ANALYSIS. J Hypertens 34 Suppl 1 - ISH 2016 Abstract Book, e539 (2016).

Ruigómez, A., Pharmacoepidemiology: an introduction. 3rd edn., (2002).

Schneeweiss, S., Avorn, J., 2005. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J. Clin. Epidemiol. 58, 323–337. https://doi.org/10.1016/j.jclinepi.2004.10.012

Ryan, P.B., Schuemie, M.J., Welebob, E., Duke, J., Valentine,S., Hartzema, A.G. Defining a reference set to support methodological research in drug safety. Drug Saf., 36 (Suppl):S33-47.

Schuemie, M.J., Hripcsak, G., Ryan, P.B., Madigan, D., Suchard, M.A., 2018. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. Proc. Natl. Acad. Sci. U. S. A. 115, 2571–2577. https://doi.org/10.1073/pnas.1708282114

Schuemie, M.J., Ryan, P.B., DuMouchel, W., Suchard, M.A., Madigan, D., 2014. Interpreting observational studies: why empirical calibration is needed to correct p-values. Stat. Med. 33, 209–218. https://doi.org/10.1002/sim.5925

Strom B L. (Editor), Hennessy S (Editor), Pharmacoepidemiology, 5th Edition. (2012), pp. 976.

van Asseldonk, D.P., Jharap, B., Kuik, D.J., de Boer, N.K.H., Westerveld, B.D., Russel, M.G.V.M., Kubben, F.J.G.M., van Bodegraven, A.A., Mulder, C.J., 2011. Prolonged thioguanine therapy is well tolerated and safe in the treatment of ulcerative colitis. Dig. Liver Dis. Off. J. Ital. Soc. Gastroenterol. Ital. Assoc. Study Liver 43, 110–115. https://doi.org/10.1016/j.dld.2010.07.004

van Wieren-de Wijer DB, Maitland-van der Zee AH, de Boer A, Stricker BH, Kroon AA, de Leeuw PW, Bozkurt O, Klungel OH. Recruitment of participants through community pharmacies for a

pharmacogenetic study of antihypertensive drug treatment. Pharm World Sci. 2009 Apr;31(2):158-64. doi:10.1007/s11096-008-9264-x. Epub 2008 Nov 30. PubMed PMID: 19043802.

Wang S. V. et al., Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0. Pharmacoepidemiology and Drug Safety 26, 1018-1032 (2017).

Wisniewski, A.F.Z., Bate, A., Bousquet, C., Brueckner, A., Candore, G., Jublin, K., Macia-Martinez, M.A., Manlik, K., Quarcoo, N., Seabroke, S., Slattery, J., Southworth, H., Thakrar, B., Tregunno, P., Van Holle, L., Kayser, M., Noren, G.N. 2016 Good Signal Detection Practices: Evidence from IMI PROTECT. Drug Saf, 39: 469-490.

4.1.16. Appendices

4.1.16.1. Appendix 1 Survey to collect additional information on the included data sources for the Big Data task force

In particular, for the <Name of data source> data source could you please provide us with the following additional information:

1. Any documentation describing data characteristics, i.e. the collected variables that allow the identification of medicines, diseases, diagnoses, laboratory data, hospital data, etc.

2. Is data collection patient-based and is the conduct of longitudinal patient-based analyses possible?

3. Population coverage and number of active patients included in the data base;

4. Is linkage to other data sources (e.g. administrative data, hospital data, death register, birth register, etc.) possible or does the data base contain comprehensive data?

5. What are the conditions (and fees if applicable) for access to the data for research purposes, e.g. if the Agency, another EU regulatory authority or an academic institution wishes either to request analyses based on the data or get access to the data for their own analyses?

6. Has the data base been subject to validation studies and if yes which algorithms were used (please provide publication(s) as applicable)?

7. Any other important information relevant for the conduct of observational research?

Appendix 2A List of initially identified data sources

Country	Data source	Acronym	Overarching data custodian	Representative population
Belgium	IMS LifeLink:Longitudinal Prescription Data (LRx) - Belgium		IMS	Uncertain
Belgium	IMS LifeLink: Hospital Disease Database - Belgium		IMS	Uncertain
EU	EUROCAT register	MFIR-Ulster	The European Commission Joint Research Centre	Yes
Denmark	Danish Civil Registration System	CPR Registry	Danish Health data protection agency	Yes
Denmark	Danish National Patient Registry	DNPR	Danish Health data agency	Yes
Denmark	The Danish Health Service Prescription Database	DHSPD	Danish Health data agency	Yes
Denmark	The Aarhus University Hospital Database	Aarhus	Aarhus University	Yes
Denmark	Danish Medical Registries (multiple)		Danish Health data agency	Yes
Denmark	Odense Pharmacoepidemiological Database	OPED	Odense University	Yes
Finland	Causes of Death Register	Cause of Death FI	Statistics Finland	Yes
Finland	Prescription Register (Finland)	Prescription Register	Social Insurance Institution	Yes
Finland	Finnish linked national health registers		National Institute for Health and Welfare	Yes
France	Securite Sociale de l'Assurance Maladie	SNIIRAM	Portail Epidemiologie France - Health Databases	Yes
France	Echantillon Généraliste de Bénéficiairs	EGB	Portail Epidemiologie	Yes

			France - Health Databases	
France	Programme médicalisé des systèmes d'informations	PMSI	Technical Hospitalisation Information Agency	Yes
France, Germany, United Kingdom	Intercontinental Marketing Services Disease Analyser	IMS	4.2 mil (UK), 29.9 mil (DE), 5.2 mil (FR)	Uncertain
Germany	German Pharmacoepidemiological Research Database	BIPS	>20 mil	Yes
Iceland	The Icelandic Medicines Registry		The Directorate of Health	Yes
Iceland	Health Service Executive Primary Care Reimbursement Services	HSE-PCRS	Health Service Executive	Yes
Italy	National drug consumption database: OsMed database	OsMeD	Agencia Italiana del Farmaco	Yes
Italy	Health Search/CSD Patient	HSD	Uncertain	Yes
Italy	Hospital Information System	HIS	Department of Epidemiology of the Regional Health Service – Lazio	Yes
Italy	Region Emilia- Romagna (RER) Database		Uncertain	Yes
Italy	ARS Tuscany database	ARS	Uncertain	Yes
Italy	Lombardia database	DENALI	Uncertain	Yes
Italy	Pedianet	Pedianet		Uncertain
Italy	Caserta database	Caserta		Yes
Lithuania	National Health Insurance Fund database		National Health Insurance Fund	Yes
Netherlands	Integrated Primary Care Information database	IPCI	Erasmus MC: University Medical Center Rotterdam	Yes
Netherlands	Agis Health Database (Achmea)	AGIS	Achmea Health Base	No

Netherlands	PHARMO Database Network	Pharmo	Pharmo	Yes
Netherlands	IMS LifeLink: Longitudinal Prescription Data - Netherlands	IMS	12 million	Uncertain
Norway	Norwegian Drug Wholesales-statistics		Norwegian Institute of Public Health	Yes
Norway	Norwegian Prescription Database	NorPD	Norwegian Institute of Public Health	Yes
Poland	Narodowy Fundusz Zdrowia (National Health Fund)	National Health Fund		Yes
Slovenia	National drug consumption database of Slovenia	Uncertain	2 million(?)	Yes
Spain	IMS LifeLink:Longitudinal Prescription Data (LRx) - Spain	IMS	3.5 million	Uncertain
Spain	Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria	BIFAP	AEMPS	Yes
Spain	The Information System for the Development of Research in Primary Care	SIDIAP database	Jordi Gol Foundation	Yes
Spain	National drug consumption database: DGFPS database	DGFPS	Ministerio de Sanidad, Servicios Sociales e Igualdad(?)	Yes
Sweden	The Swedish Prescribed Drug Register		Swedish National Board of Health and Welfare	Yes
Sweden	Swedish Medical Birth Register		Swedish National Board of Health and Welfare	Yes
Sweden	Swedish National Patient Register		Swedish National Board of Health and Welfare	Yes

UK	Clinical Practice Research Datalink - Primary care	CPRD	CPRD	Yes
UK	The Health Improvement Network - Primary care	THIN	INPS	Yes
UK	QRESEARCH		Qresearch	Yes
UK	The electronic Data Research and Innovation Service	eDRIS	Information Services Division Scotland	Yes
UK	Secure Anonymised Information Linkage	SAIL	University of Swansea/Welsh Government	Yes

Appendix 2B List of data sources retained for further characterisation

	Data source name	Country	Type	Type of care	Start date
1	QuintilesIMS LifeLink: Hospital Disease Database – Belgium	Belgium	Electronic healthcare records	Secondary care	2001
2	Danish National and regional registries [1]	Denmark	Record linkage system	Mixed	1977
3	Finnish National registries [2]	Finland	Record linkage system	Mixed	1964
4	Securite Sociale de l'Assurance Maladie (SNIIRAM)	France	Claims	Mixed	1999
5	Echantillon Généraliste de Bénéficiaires (EGB)	France	Claims	Mixed	2006
6	QuintilesIMS Disease Analyser	France	Electronic medical records	Primary care	1997
7	QuintilesIMS Disease Analyser	Germany	Electronic medical records	Mixed	1992
8	German Pharmacoepidemiological Research Database	Germany	Claims	Mixed	2004
9	Icelandic Registries[3]	Iceland	Record linkage system	Mixed	Unk
10	Pedianet Database	Italy	Electronic medical records	Primary care	1998
11	Agencia Regionale di Sanita Tuscany database	Italy	Claims	Secondary care	1996
12	Hospital Information System –Lazio Region	Italy	Electronic medical records	Secondary care	Unk
13	Lombardia Health Database	Italy	Claims	Secondary care	2000
14	QuintilesIMS LPD Health Search Database Longitudinal	Italy	Electronic medical records	Primary care	2000
15	Region Emilia Romagna Database	Italy	Claims	Secondary care	Unk
16	Integrated Primary Care Information Database	Netherlands	Electronic medical records	Primary care	1995

17	VEKTIS	Netherlands	Claims	Mixed	Unk
18	Pharmo Database Network	Netherlands	Record linkage system	Mixed	1990
19	Norwegian Registries [4]	Norway	Record linkage system	Mixed	1997
20	National Health Fund	Poland	Claims	MIxed	Unk
21	The Information System for the Development of Research in Primary Care	Spain	Electronic medical records	Primary care	2006
22	Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria	Spain	Electronic medical records	Primary care	2002
23	Information System of Parc de Salut del Mar	Spain	Electronic medical records	Secondary care	Unk
24	QuintilesIMS LPD Health Search Database Longitudinal	Spain	Electronic medical records	Mixed	2006
25	Swedish National Registries[5]	Sweden	Record linkage system	Mixed	1970
26	Clinical Practice Research Datalink - Primary care	United Kingdom	Electronic medical records	Primary care	1987
27	QResearch	United Kingdom	Electronic medical records	Primary care	NA
28	The electronic Data Research and Innovation Service	United Kingdom	Record linkage system	MIxed	Unk
29	Secure Anonymised Information Linkage	United Kingdom	Record linkage system	Mixed	Unk
30	Hospital Treatment Insights	United Kingdom	Record linkage system	Secondary care	2010
31	The Health Improvement Network - Primary care	United Kingdom	Electronic medical records	Primary care	2003
32	QuintilesIMS LPD Health Search Database Longitudinal	France	Electronic medical records	Mixed	
33	Caserta Database	Italy	Claims	Primary care	2002
34	Medicines Monitoring Unit Scotland	Scotland	Record linkage system	Mixed	1990

1 Danish Civil Registration System (CRS) + Danish National Patient Registry (DNPR)-HOSPITAL based + Odense Pharmacoepidemiological Database+The Danish National Database of Reimbursed Prescriptions + Phepi prescription database in Northern Denmark Database of the Central Denmark Region Health Services (Aarhus))

2 Causes of Death Register Finland + Finnish Linked National Health Registers + Finnish Prescription Register +Medical Birth Register +National Hospital Discharge Register +Register for Congenital Malformations +Register for Induced Abortions + Register of Primary Health Care Visits)

3 Norwegian Drug Wholesales Statistics Norwegian Prescription Database Norwegian Hip Fracture Register Norwegian Medical Birth Register Norwegian Registry of Pregnancy Termination the Cause of Death Register Cancer Registry Norwegian Patient Registry (hospital discharge registry) Norwegian Cardiovascular disease registry)

4 National Patient Register (NPR) + Swedish Cancer Register +Swedish Cause of Death Register+ The Swedish Prescribed Drug Register

5 The Icelandic Medicines Registry the Icelandic National Patient Registry, Registry for Causes of Death

4.1.16.2. Appendix 3 Dashboard created for visual representation of the scoring of the included data sources

https://www.ema.europa.eu/documents/report/dashboard-created-visual-representation-scoring-included-data-sources_en.xlsx

4.2. *Registry data*

4.2.1. Background

As defined by the EMA, patient registries are organised systems that use observational methods to collect uniform data on a population defined by a particular disease, condition, or exposure, and that is followed over time (EMA website). While data from randomised clinical trials typically provide the primary evidence for marketing authorisation of a new product, potential risks and benefits are not fully known at this stage. Patient registries contain information of value for filling these evidence gaps and thereby are a central source of big data in the area of healthcare (EMA website; EMA report: "Patient Registries Workshop", 2016). For example, registry data can be used during the development phase of a new product to define target populations and to identify unmet needs and estimate the disease burden. During the marketing authorisation evaluation procedure, these data can provide historical controls. Finally, in the post-marketing phase, registry studies are performed to assess safety and effectiveness and to investigate off-label use.

Although registry data have many applications, the data quality, completeness and the possibilities of linking data to external data, sources (such as e.g. prescription data) are highly variable across different registries, providers and countries.

The number of initiatives to improve, coordinate and harmonise patient registries have increased during the recent years. In addition, the need to pool data across different databases has led to an increase in methods for pooling data to obtain larger, stronger and more useful data sets. As an example, the recent RD-Connect project on rare diseases has established an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research (EMA report: "Patient Registries Workshop", 2016).

In Europe, the EMA launched a registry initiative in 2015, aiming to facilitate the use of registries to better support the authorisation of medicines (EMA: Initiative for patient registries – Strategy and pilot phase, 2015). The initiative is mapping ongoing projects at national and international levels and aims to provide guidance regarding standardised methodological approaches when creating a new registry (EMA report: "Patient Registries Workshop" 2016). The scope of the EMA registry initiative overlaps with that of the HMA/EMA Big Data Task Force subgroup for observational data.

4.2.2. Objectives

The objectives of this report are twofold. First, we present the characteristics of existing European registries and discuss main data formats and access to data for the purpose of conducting population-based observational studies. Second, we present a list of applications of registry data within the regulatory process. Advantages and disadvantages are presented as well as several examples of application of registry data throughout the life cycle of a product.

Not included in the scope

Observational data from clinical trials (non-interventional studies) are not included; this data source is covered by the Clinical Trial Subgroup of the Big Data Task Force.

4.2.3. Methods

Information for the mapping and characterisation of registries has been obtained from the following sources:

a. The EMA Registry Initiative. Under this initiative, the EMA provides a platform where interested parties can find information about different healthcare databases, networks and research organisations in the EU (EMA website). It is possible to perform a search by centre, network or database. In particular, we used the data sources inventory (Gross list of registries, Excel spreadsheet, which is attached to this report).

b. Literature search on relevant registries.

We will discuss six registries in detail representing European registries, recognising that it would be too demanding to map and characterise all relevant data sources in detail. These registries were selected based on an expert judgement since they hold high-quality longitudinal data, are based in different European regions and cover diverse disease areas and applications. In addition, they are suited for external collaboration and linkage with other databases.

In order to describe the applicability of patient registry data in the regulatory process, a thorough literature search was also carried out.

Characterisation of patient registries

The following aspects are described for each of the selected registries:

- Data structure, provenance of data and updates.
- Data quality: validity and completeness.
- Accessibility of data, methodology and data linkage possibilities.

Challenges related to the applicability of registry data in the regulatory context are also considered.

Core data elements/sets

A definition of a core, minimal data set (core data set) within the type of registry (disease) is suggested. The EMA Registry Initiative highlighted the need for regulators to provide guidance about the minimum level of data and quality parameters required for the applicability of the registries in regulatory decisions (EMA report: "Patient Registries Workshop" 2016).

4.2.4. Key case studies

4.2.4.1. General considerations related to characterisation of registries

Six registries have been selected for detailed description and characterisation (Appendix I.1 – I.6).

- Appendix I.1: Netherlands Cancer Registry (NCR).
- Appendix I.2: Danish National Health Registries.
- Appendix I.3: European Society for Blood and Marrow Transplantation (EBMT).
- Appendix I.4: European Cystic Fibrosis Society (ECFS).
- Appendix I.5: European Registry for Multiple Sclerosis (EUReMS).
- Appendix I.6: British Society for Rheumatology Biologics Registries (BSRBR).

4.2.5. Data characterisation

4.2.5.1. Data structure, provenance of data and updates

The data in all six registries are structured, although not in the same way. No overall standard and harmonisation are applied across the registries. In general, variables are clearly defined within each registry and have finite number of possible values. Data are either collected electronically or manually from patient records and hospital information systems (NCR and DNPR). The data are either entered in a central database (EBMT, ECFS, BSRBR) or collected from several registries across countries (EUReMS). Data may be uploaded continuously, on a monthly basis (DNPR) or yearly (ECFS).

Data are collected at patient level, and access to these data is restricted and regulated.

4.2.5.2. Data quality: validity and completeness

Overall, no data-quality standards or rules exist, neither in regard to our selection of registries nor registries in general. In this context, data of sufficient quality are defined as data that are suitable for the intended purpose of analysis, which does not necessarily imply that the database is correct and complete. For example, a registry may not be useful to evaluate efficacy during the pre-authorisation process but could be extremely valuable to identify off-label use and potential side effects, in particular in combination with other medication.

The data in the selected registries seem to be accurate, and accuracy checks (NCR, DNPR, EBMT, ECFSPR, EUReMS) are regularly performed by internally trained data managers. However, since data content and definitions of variables may change over time (e.g. for the DNPR), it is necessary to have in-depth knowledge of the implemented changes when using and analysing data. In the case of the DNPR registry, specific documentation and information on such changes are continuously published and available to users. Systematic reviews of completeness and validation of variables are performed regularly (NCR, DNPR, EBMT, ECFSPR). In the case of the ECFSPR registry, automatic controls and validation rules are applied at data entry level. It is important to keep in mind that these six registries were selected based on their high quality of data. Other registries may lack data control validation processes and may hold large amounts of incomplete data.

4.2.5.3. Accessibility of data, methodology and data linkage possibilities

Data in aggregated or summary form are available for all of the selected registries. Some of the registries publish summary reports on their websites (NCR, DNPR, ECFS). Access to data at patient level is subjected to data protection legislation. None of the selected registries can be accessed directly by the pharmaceutical industry. Access to anonymous data at patient level is possible for the DNPR, EBMT, ECFS and BSRBR registries. Guidelines to request data are available at the registries' websites, describing how and where researchers can apply for access to data for specific projects. Scientific committees get involved to ensure that the data are used according to legislation. In the case of other registries, guidelines to access data may not be available.

In regard to the EUReMS registry, all data providers retain full ownership of contributed data, including the right to withdraw it. Data can be shared with researchers and policymakers who wish to participate in the EUReMS studies' platform.

In general, the Regulatory authorities may access to data by establishing collaboration with different registries. For instance, the NCR in collaboration with NCCO (Netherland Comprehensive Cancer Organisation).

Several of the selected registries (NCR, DNPR, BSRBR) enable the linkage of data from other data sources. For example, data on drug consumption is extremely important in several effectiveness and surveillance studies and this information is rarely stored in disease registries. Thus, linkage between disease registries and drug registries are regularly performed.

Data from the NCR registry can be linked to for example data in the PHARMO database (Institute for Drug Outcomes Research). The linking process requires that each patient has a "key" or unique identification number. Linking can be also performed in cases where keys are not available, but this process is very complicated.

In the case of the BSRBR registry, data are linked to other NHS databases, such as the UK cancer and death registries. Data from the DNPR registry can be linked to other Danish registries and data sources, using the unique civil registration number (CPR), as further described in Appendix I.2. Since other Nordic countries have also implemented a civil registration number, Nordic registries are perfectly suited for linkage.

For the EBMT registry, no unique person-specific identifier is available yet.

4.2.5.4. Need for data standards

Terminologies

Based on the recommendations from the EMA Registry Initiative and the characterisation of the registries of this report, standardisation and harmonisation are needed within each disease area with respect to data collection and coding. In the long term, the main goal is to standardise terminology and variables across diseases. Therefore, the use of international classifications standards for coding of for example diagnosis is highly encouraged.

International standards for clinical terminologies such as SNOWMED CT (https://www.opencimi.org/tag/SNOWMED%20CT) could be implemented. SNOWMED CT is a systematically organised computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. Another international terminology standard is the International Classification of Diseases (ICD) (http://www.who.int/classifications/icd/en/), developed by the WHO. It is a standard diagnostic tool for epidemiology, health management and clinical purposes. MedDRA (Medical Dictionary for Regulatory Activities) is a clinically validated international medical terminology dictionary implemented by regulatory authorities in the pharmaceutical industry. MedDRA is used during the regulatory process, from pre-marketing to post-marketing activities, and for data entry, retrieval, evaluation and presentation. MedDRA is the adverse event classification dictionary endorsed by the ICH (International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use). The Anatomical Therapeutic Chemical (ATC) Classification System, controlled by WHO (http://www.who.int/medicines/regulation/medicines-safety/toolkit_atc/en/), is generally used for the coding of active drug ingredients.

Data integration

There is a need for data integration across different registries, observational sources and other types of data. Several examples of collaborations between registries – networks of registries – are listed below.

- The EUReMS project is an initiative of the European Multiple Sclerosis Platform. The project identifies and pools multiple sclerosis (MS)-related data from different regions. Twelve registries

have started pooling their data according to an agreed protocol to harmonise heterogeneous MS information (Appendix I.5).

- In the field of cancer, the European Network of Cancer Registries (ENCR, http://www.encr.eu/) encourages collaboration between population-based cancer registries by coordinating activities and mapping of priorities for research topics (Appendix I.1).
- The European Bone Marrow Transplantation (EBMT) registry is the single biggest data source of its kind in Europe, covering more than 500 centres in approximately 50 countries. The registry provides a number of services to registry users, e.g. a scientific and educational programme and market surveillance in collaboration with health authorities (Appendix I.3).
- The European Cystic Fibrosis Society (ECFS) patient registry includes demographic and clinical data of cystic fibrosis patients from 27 countries. Data are collected using the data-collection platform ECFSTracker, an open-source multipurpose and multinational software program. The aim is to measure, survey and compare aspects of cystic fibrosis and its treatment across countries to encourage new standards of dealing with the disease (Appendix I.4).
- The Danish National Health Registries are included in European database networks with data from other Nordic countries. Different models of data networking are applied. (Appendix I.2).

There are differences between routine records accumulation in systems like Mini-Sentinel or Observational Medical Outcomes Partnership (OMOP, US) and European networks of registries. In many European countries there are linkage opportunities among several health care databases that allow following the subjects during their entire life (Ehrenstein et al. 2017).

Technical, logistical, ethical and legal challenges affect the assembling of such database networks, and they are difficult to overcome. Practical guides of the different models of data networking have been provided by Ehrenstein et al. (Ehrenstein et al. 2017) and Gini et al. (Gini et al. 2016). Usually, a global protocol is followed. Depending on the aims of the study, the analysis can be based on limited sharing or sharing involving the harmonisation and pooling of individual data in a common data model (CDM), whereby partners transform data to create standard input data sets according to specifications (Ehrenstein et al. 2017).

The integration of several data sources is of great advantage in the area of epidemiology and pharmacoepidemiology. For example, increased precision of results can be obtained, enhancement of small potential risk signals related to newly marketed therapies is facilitated, and there are better possibilities to investigate rare adverse events and infrequent exposures. But there are several challenges as well; the usual epidemiological challenges related to validity are not addressed by big data. Indeed, validity concerns can increase when several databases are combined. Large amounts of missing data may cause selection bias, and the well-known challenges of applying observational data such as reverse causation, immortal time bias, and healthy user/healthy adherer bias are not remedied by large amounts of data (Ehrenstein et al. 2017).

Depending on the registry, a huge amount of work may be required to link data across registries. A high degree of standardisation and harmonisation will ease this process and improve feasibility. When it comes to linking registry data with other data sources at the individual patient level, the presence of a personal identification number (PIN) is of paramount importance.

4.2.6. Applicability of registries in the regulatory process

The life cycle of a pharmaceutical product can be divided in three phases: development, authorisation and post-authorisation. Registry data can provide further insights in all those phases (see Table 1). For example, during drug development, registries can be used to gather information about the target

population, current standard care and the epidemiology of the disease. In the case of rare diseases, the European Commission has recognised that patient registries are key instruments for improving healthcare planning and clinical research (European Commission, 2017).

During the authorisation phase, information about the safety and efficacy of the drug is gathered. Randomised controlled clinical trials (RCTs) are the gold standard for the approval of new indications. However, generalising their results to a broader population has constraints; trials are usually conducted in narrowly defined populations, follow a carefully designed protocol, and data are collected over a few years only. In some cases, a randomised controlled trial may not even be feasible or appropriate. Therefore, data obtained from other sources can provide valuable insights during the evaluation of the safety and efficacy of new products (Spigel D. R, 2010).

Hastwell et al. performed a review of FDA and EMA approvals of pharmaceuticals between 1999 and 2014 (Hastwell et al, 2016). During this period, the EMA granted 795 new applications, of which 44 were new indications for 35 drugs without an RCT. It was mainly within the oncology disease area that uncontrolled evidence was accepted and mostly within haematological malignancies and solid tumours. In the majority of these applications, the main studies did not include a direct comparison between effect of the product and other medicines or placebo. For example, the effect of cyanokit (hydroxocobalamin) was evaluated by studying records of treated patients (Cyanokit, EPAR). In other cases, registry data were used in the main studies to select historical controls. During the investigation of alglucosidase alfa (Myozyme, EPAR), a historical cohort was used as a comparison group since it was unethical to include a placebo group (Kishnani et al, 2007). Eculizumab (Soliris) was initially approved in 2007 for the treatment of paroxysmal nocturnal haemoglobinuria (PNH). In 2016, the EMA granted an extension of the label to other indications using historical controls from the PNH Registry (Soliris, EPAR).

After approval, registry data are collected for safety and effectiveness studies because the knowledge of benefits and risks is limited at the time of approval (Suvarna, 2010). A retrospective analysis of the central procedures for marketing authorisation during the years 2005-2013 shows that the EMA requested the creation of registries or use of existing registries for further studies in approximately 7% of cases (Bouvy et al, 2017). Approximately 70% of the studies had a primary safety objective.

Phases in the pharmaceutical life cycle		Applications of registry data
Development	Orphan designation	Description and quantification of the target population
	Paediatric investigations	Description of the standard care received by children
	Evaluation of unmet medical need	Studies based on disease registries enable descriptions of the current standard care
Authorisation	Efficacy studies	Use of historical controls
		Demonstration of surrogate endpoints adequacy
	Effectiveness studies	Comparisons between the new drug and other treatments
		Subpopulation studies

		Studies with multiple endpoints
	Patient-reported outcomes (PROM) studies	PROM studies across different populations and countries
	Biomarkers	Validation of proposed biomarkers and discovery of new ones
Post-authorisation	Label extension	Off-label use studies
	Safety studies	Longer follow-up to find unknown adverse reactions
		Drug-drug interaction studies

Table 1: Applications of registry studies in the different phases of the product life cycle.

4.2.7. Regulatory acceptability of registries in the regulatory process

While registry-based studies add valuable information during the life cycle of a medicine, they have several limitations. Patient registries are not created to answer a particular research question; they collect information about the treatment given by local physicians according to what they know. In real life, treatments are not administered randomly to patients, and a line of variables and other factors are not measured or monitored. This makes it difficult to correct for potential bias during the analyses.

In addition, since registry data and RCT data are not directly comparable, different results can be obtained when studying the same question (Avorn J., 2007). The implementation of registry data within the regulatory context is further impeded by data protection issues and the lack of access to data.

4.2.8. Solutions for improving regulatory acceptability

Patient registries are already an important tool in the evaluation of medicines, in particular during the post-approval phase. However, in order to further increase their role during the product life cycle, several actions could be implemented.

4.2.9. Standardisation of registries

It is recommended to standardise existing and future registries. Therefore, we propose to use a core, minimal data set within any type of (disease) registry. The following information should be included in the core data set:

- Demographics:
  - Centre code,
  - Patient code,
  - Date of birth,
  - Gender,
  - Cause of death,
  - Date of death.
- Diagnosis:

  - Diagnosis (coded in accordance with an international standard),
  - Date of diagnosis confirmed (to estimate time since first diagnosis).
- Therapy/treatment:
  - Start/stop date,
  - Generic name, concomitant medications.
- Complications/co-morbidities:
  - Diagnosis (coded in accordance with an international standard),
  - Date of diagnosis.
- Genotype (when applicable).
- Follow-up.

4.2.10. Recommendations from the EMA Registry Initiative

In 2015, the EMA launched the 'Initiative for patient registries' to make better use of existing registries and to facilitate the establishment of new ones. The outcomes of this initiative are very important for the further implementation of registry studies in the regulatory context. The EMA Registry Initiative held a workshop in 2016 to explore the perspectives of multiple stakeholders, including registry holders, patients, the pharmaceutical industry, health technology assessment representatives and regulators (EMA Registry Initiative).

A comprehensive list of recommendations within the below five theme areas was an important outcome of this workshop:

Benefits of patient registries and obstacles to be overcome

- To facilitate stakeholder collaborations, incentives are needed for registry holders to collect data to meet needs that are not directly their own.
- Technical challenges could be overcome through standardised data collection, coding, and analytical procedures and by linking registry data to external data sources.
- Appropriate governance procedures should be developed to safeguard transparency, accessibility of data and the independence of registries, and to provide clarity about legal and regulatory requirements relating to patient registries.

Benefits and challenges of collaboration

- Studies potentially involving registries, including post-authorisation studies, should be planned early in the product development phase. Stakeholders, including regulators, should communicate directly with each other in the planning of studies to agree on outcomes and recognise limitations.

Technical considerations

- Regulators should provide guidance to registry holders on the core data elements and quality parameters considered to represent an acceptable standard to support regulatory decision-making.
- Data collection, quality and interoperability should be improved through use of standardised data fields, dictionaries and coding. The proposed core data set presented here is very much in line with the core data set proposed by the EMA Registry Initiative.

- Technological advances should be exploited to increase patient participation and to improve the value of registries in clinical care by facilitating linkage with other healthcare datasets, data pooling and analyses.

Governance

- Consents obtained from patients should be clear about data sharing and access for stakeholders other than the registry holders with appropriate consents in place for levels of data sharing.
- Principles of governance should be established to guide interactions between registries, pharmaceutical industry and regulators addressing data privacy, ownership, financial aspects, transparency, commercial-in-confidence agreements, and accessibility of data for public health purposes.

Sustainability

- Sustainability should be based on a development model, a professional management structure and the development of clear partnership with stakeholders to safeguard independence.

4.2.11. Data quality and data protection

Since registry data are collected for other purposes, it is a challenge to use them in clinical studies. One of the highest priorities is to characterise and improve the data quality of patient registries. From a methodological point of view, statistical models can be used to partially compensate for confounding factors (Freemantle et al, 2013). However, since the choice of model will affect the conclusions to a greater extent in observational studies, regulatory guidance on the implementation of statistical methods is needed. It will be necessary to align policies on the use of real-world data to promote its use in marketing applications and for regulatory issues such as safety.

The data stored in patient registries are highly sensitive. Patient consent and data protection issues could be addressed through the implementation of a European standard procedure for registry studies.

4.2.12. Increase collaboration between stakeholders

There are several stakeholders involved in registry studies. Patients need to agree to participate in the registry. The registries have an administrative structure and must comply with data protection regulations. Pharmaceutical companies require access to data to assess their products. Academia carries out epidemiological studies regularly. Authorities also require access to data to investigate safety issues. Therefore, collaboration between patient organisations, pharmaceutical companies, academia and regulators is of paramount importance.

4.2.13. Conclusions

This report discusses the main characteristics of patient registries and the implementation of registries in the life cycle of a medicine.

The selected examples of registries described in Appendix I illustrate initiatives and registries of high quality and integration potential. They are model registries and may inspire others to optimise their use of registries for regulatory decision-making.

In general, it seems that registries are not sufficiently coordinated, neither domestically or across borders, and there is a lack of awareness of the existing registries. In addition, the registries appear to be lacking harmonised protocols, scientific methods and data structures across registries. Limited data sharing between registries is frequently observed despite the fact that data sharing offers multiple

benefits such as increasing cohort size, powering studies and finding confirmatory cases. In addition, there is uncertainty regarding what data are needed and what data are being collected.

Even though patient registries are already used in the regulatory process, their full implementation is limited by insufficient regulatory guidance regarding accepted areas of applications and methodological practice. The EMA Patient Registries Initiative aimed to explore ways of expanding the use of patient registries. Some actions taken by this initiative included support to the interaction between patient registries, industry and academia; and exploring the possibility for a qualification process for registries. However, further work is needed.

Other factors limiting the use of register data are related to data protection and patients' informed consent. The recently implemented General Data Protection Regulation should be followed when conducting registry studies and the impact of this regulation is yet to be seen.

Based on the presented analysis, the following actions are recommended:

- Use of standardised data fields, dictionaries and coding should be implemented to improve data collection, quality and data interoperability.
- The sharing of information between registries within a disease area should be encouraged.
- Provision of guidance on governance principles and standards for transparency, accessibility and stakeholder interaction.
- Definition of core data elements.
- Provision of methodological and technical guidance regarding data collection – both for existing registries and newly established ones.
- Facilitation of access to data for different stakeholders.
- Provision of guidance on accepted methods in registry-based studies with different purposes, e.g. safety, efficacy, effectiveness.
- Establishment of a European standard for registry studies concerning data protection and patient consent.
- Alignment of policies on the acceptance of real-word data across different regulatory bodies in Europe.

4.2.14. Appendix I – Examples: Characterisation of selected data sources – registries and platforms

4.2.14.1. Appendix I.1.: Netherlands Cancer Registry (NCR)

Field: Cancer
Disease: Cancer
Based in: Netherlands
Website: https://www.iknl.nl/over-iknl/about-iknl/what/. Input has also been obtained from employees of the Netherlands Comprehensive Cancer Organisation (IKNL).

Background

Information about every patient with cancer is gathered in the Netherlands Cancer Registry (NCR) by the Netherlands Comprehensive Cancer Organisation (IKNL). IKNL is a national organisation in the Netherlands that also plays an independent role in regional network and supports collaboration in oncological care. The IKNL combines 8 regional oncologic care centres. From 1989, data have been collected about patients diagnosed with cancer in the NCR. Initially, only the primary treatment was registered.

Since 2014, additional details, including subsequent treatments, have been collected for a number of cancer diseases. Due to the extended data set, the NCR is even more relevant for healthcare professionals, healthcare institutions and researchers. The goals of the NCR are to improve quality of care, and ultimately to improve survival in patients with cancer.

It includes:

- About 17 million inhabitants of the Netherlands, 80 hospitals.
- Annually, 100,000 new cancer diagnosis, and 44,000 cancer deaths in the Netherlands.
- All patients are included with a pathological confirmation of cancer, except when a patient objects (only 1-2 patients a year). The NCR has an opt-out system.

The data in NCR are reported in three domains: the public domain (science), the political domain (Ministry of Health, Welfare and Sport; National Health Care Institute), and the care domain (hospitals/care institutions, professionals and patients). Information is used to support individual hospitals and care institutions in policy-making. The IKNL responds to developments in the field by shifting its focus from institutional to transmural and regional and also from a general to a tumour-specific approach to oncological care (www.iknl.nl).

Collaboration with MEB and IKNL

Last year, the Medicines Evaluation Board of the Netherlands (MEB) collaborated with the IKNL. In centralised procedures for marketing authorisation of new medicinal products or applications for new indications, questions have been posed to the IKNL about data from the NCR to optimise the assessment of the centralised procedure. A total of five questions were submitted. Two of these questions could be answered and were also brought up for discussion in the CHMP. The remaining three questions could not be answered because the information requested had to do with second-line or third-line treatment, for which data regarding medicinal products are not yet available. During the coming years, this information will be collected in the NCR (www.iknl.nl).

Other cancer registries in Europe

The NCR is one of the 160 population-based cancer registries (CRs) in Europe (Coebergh, J.W. et al. 2015). Most cancer registries cover all cancers, but some are confined to specific cancers or to children. They cover 15–55% of the populations in all of the larger member states of the European Union (EU), except the United Kingdom (UK), and 100% coverage in 80% of those with populations below 20 million. The potential of CRs for clinical evaluation has grown substantially through interaction with clinical stakeholders and more incidentally biobanks, also with greater involvement of patient groups – with a special focus on elderly patients who generally do not take part in clinical trials. Whereas 25-35% of CRs are active in a range of cancer research areas, the rest have a low profile and usually provide only incidence and survival data. The perception of unity in diversity and suboptimal comparability in performance and governance of CRs was confirmed in the EUROCOURSE (EUROpe against cancer: Optimisation of the Use of Registries for Scientific Excellence in research) European Research Area (ERA)-net coordination FP7 project of the European Commission (EU) which explored best practices, bottlenecks and future challenges of CRs (Coebergh, J.W. et al. 2015; www.eurocourse.org). Despite access to specialised care-related shortcomings, especially of survival cohort studies, European databases for studies of incidence and survival (such as ACCIS and EUREG on the one hand and EUROCARE and RARECARE on the other hand) have proved to be powerful means for comparative national or regional cancer surveillance (www.iknl.nl).

Characterisation (NCR)

Data structure, provenance of data and updates

The data collected in the NCR is structured. Tumour-specific item sets are created and collected for all patients who are diagnosed with cancer. Data are collected from patient records and hospital information systems.

Data quality: validity and completeness

- Data managers follow an internal one-year training course and receive continuous training to keep them up to date.
- National and international coding guidelines are used (i.e. for CTC AEs).
- Database with many quality checks.
- Data and results of analyses are discussed with partners in the healthcare field.
- International comparison with other registries to check validity.

Accessibility of data, methodology and data linkage possibilities

The NCR is a national registry. It identifies patients through several sources (pathology laboratories (+/-95%), Landelijke Basisregistratie Ziekenhuiszorg (LBZ), DBCs (+/-5%). Completeness is different for different tumours, but generally believed to be at least 95%.

It is also possible to combine NCR data with other data sources, for example the PHARMO database (Institute for Drug Outcomes Research, http://pharmo.nl). The social security number is not collected in the NCR, which sometimes makes it more difficult to link data. It is possible to obtain anonymous

patient-level data, but most data requests require aggregated data.
Regarding the accessibility of data for NCAs, some of the data is published on http://www.dutchcancerfigures.nl. Data on incidence, prevalence, survival, mortality and risk are included on this website. Data can be accessed in collaboration with the IKNL.

EUROCOURSE project
The major results from the EUROCOURSE project by work package (WP) can be found on the www.eurocourse.org website.

- WP 1.3 Survey on research and funding
- WP 1.4 Best CR practices
- WP 1.5 In search of programme owners
- WP 1.6 Governance for programme owners
- WP 2.2 Confidentiality guidelines
- WP 3.3 Data quality control
- WP 4 Exploration of potential users by research domain
- WP 4.5 European Cancer Observatory
- WP 5 Guidelines for linkage of CRs to screening registries
- WP 6.3 State of the art of effective use of registry indicators in evaluating cancer care
- WP 6.5 Overview of clinical cancer registries in Europe
- WP 7 Guidelines on linkage between biobanks and CRs
- WP 8 International collaborative studies by research domain
- WP 9.2 Report of Cancer Registry Summit at ECCO Oncopolicy Meeting
- WP 9.3 Brochure on CRs in Europe and role of European Network of Cancer Registries

4.2.14.2. Appendix I.2.: Danish National Health Registries

Field: Family, hospital, and disease registries
Disease: All diseases.
Based in: Denmark
Website: https://www.sst.dk/en

Background

In Denmark (and the other Nordic countries), government-funded universal health care in combination with a tradition of keeping records enabling linkage at the individual level have led to the establishment of extensive networks of inter-linkable and longitudinal population-based registries covering entire nations. Patient registries covering entire nations with individual-level linkage potential have existed in Denmark since 1968, Finland since 1969, Sweden since 1987, Iceland since 1999, and Norway since 2008.

The Danish healthcare system

Inhabitants of Denmark as of May 2017: 5,714,910 people – excluding inhabitants of Greenland and the Faroe Islands.

The Danish healthcare system can generally be described as having three administrative levels. The first level is represented by the state, which is responsible for legislation, national guidelines, surveillance and health financing through the Ministry of Health. The second level is represented by the five regions, which are responsible for the provision of primary and hospital care. The third level is represented by the 98 municipalities, which are responsible for a broad range of welfare services such as school health, child dental treatment, home care, primary disease prevention and rehabilitation.

The Danish National Health Service provides tax-supported healthcare for the entire Danish population (Schmidt, M. et al. 2015).

The National registries

The nationwide registry of administrative information, the Danish Civil Registration System, was established in April 1968 and is one of the oldest in Europe. It is a key tool for epidemiological research in Denmark. All persons residing in Denmark are assigned a unique ten-digit personal identification number, the Civil Personal Registration (CPR) number. It allows for technically easy and exact linkage of Danish registries at the individual level.

Denmark thus has a long tradition of creating nationwide administrative and health registries that enable linkage at the individual level. Examples include the Danish registries on causes of death, hospitalisations and cancer, and on socioeconomic parameters such as income and education. The registries that are important for epidemiological and pharmacoepidemiological research are the Danish National Patient Registry (DNPR) established in 1977 (Schmidt, M. et al. 2015) as well as the National Prescription Registry established in 1994 (Pottegård, A. et al. 2015), which includes individual-level data on prescriptions filled by Danish residents at community pharmacies.

Figure 1 shows the timeline for the initiation of selected Danish registries linkable to the DNPR by calendar year (Schmidt, M. et al. 2015). It illustrates the considerable potential of cross-linking various administrative and clinical registries in Denmark, using the CPR number. The registers in Denmark are

numerous and comprehensive – even when benchmarking against the high standards of the Nordic countries.

Figure 1: Timeline for the initiation of selected Danish registries linkable to the Danish National Patient Registry (Schmidt, M. et al. 2015).

Network – multinational initiatives and studies using Danish National Health Registries

Recent years have seen the emergence and establishment of "big data" or networks of collaborations in epidemiology and pharmacoepidemiology mainly among the Nordics countries, continuing a long tradition of using registry data for medical research. There are several examples of successful collaborations between different registries in the Nordic countries. For example, the Nordic Arthroplasty Registry Association (NARA) studies the suitability of different types of hip replacement surgeries using data from Norway, Denmark, Sweden and Iceland.

European database networks have been established, including the ones encompassing the Nordic data, and have found ways to overcome challenges with respect to differences in the underlying healthcare systems, languages, data-sharing laws, record-generating mechanisms and classifications (Ehrenstein et al. 2017). Medical data in the Nordic countries are coded using a common basic set of standard classifications.

It has proven possible to map Nordic registry data to a common data model, the Nordic Common data Model (NDM), as part of the Caring project (Cancer Risk and Insulin Analogues) (Andersen et al. 2015; www.caring-diabetes.eu). The mapping provides sufficient power to investigate rare adverse events and infrequent exposures. The aim of the Caring project was to obtain precise data on the incidence of cancer in diabetic patients and determine any link with use of various insulin and insulin analogues. The study utilised high quality prescription databases and other national data sources, integrated at European level with advanced methods of harmonising data, and took potential confounders into account. The project aimed to determine the influence of drug dose on risk, and, through a risk model, identify predictors of cancer for insulin users http://www.caring-diabetes.eu/?q=content/general-introduction-0.

Biobanks

Biological specimens from large population groups coupled with detailed phenotypic information provide unique opportunities for genetic epidemiology studies and can also give researchers precise information, e.g. on environmental, nutritional or pharmacological exposures in the population. Statens Serum Institut under the auspices of the Danish Ministry of Health has long performed epidemiological research taking advantage of the resources available in e.g. the DNBC biobank and the Danish Newborn Screening Biobank, which is hosted by Statens Serum Institut and contains dried blood spot samples for virtually all Danes born since 1982 http://www.ssi.dk/English/Service/AboutSSI.aspx.

To further strengthen research opportunities, the Danish National Biobank was established at Statens Serum Institut. Inaugurated in March 2012, the biobank boasts state-of-the art freezers, robotic systems and laboratory facilities, and is planned to contain 15 million biological specimens collected in the Danish healthcare system. In addition to the physical biobank, the Danish National Biobank also includes an on-line biobank registry, which links information about available biological specimens with disease codes and demographic information from national registries. A search in the registry makes it possible to look up the number of biological specimens available for patients with a certain diagnosis.

The Danish National Patient Registry (DNPR)

The aims of collecting data in this registry are to monitor the frequency of various diseases and treatments, to provide a sampling frame for longitudinal population-based and clinical research, to facilitate quality assurance in Danish healthcare services as well as to facilitate hospital physicians' access to patients' hospitalisation histories.
The DNPR was established in 1977. The registry collects data from hospitals: admissions and in

addition contacts to emergency rooms and outpatient clinics since 1995. The registry does not contain information on primary care.

At the start, the registry included information on inpatients in somatic wards. The registry has since been gradually expanded, and from 2007 the DNPR has included information on all patients in Danish hospitals. The DNPR is a unique data source, however, researchers using the data should carefully consider potential fallacies in the data before drawing conclusions (Sørensen et al. 2009; Schmidt et al. 2015).

Characterisation (DNPR)

Data structure, provenance of data and updates

The registry contains structured data with each variable having a finite number of possible values. There are several variables, including administrative data (personal and admission data), civil registration number (CPR number), dates of admission and discharge, hospital and department data, diagnosis codes and surgical procedures. Information reported to the DNPR includes administrative data, diagnoses, treatments and examination.
Diagnoses are coded according to the International Classification Diseases (ICD) from 1977-1993 (ICD-8) and according to ICD-10 since 1994. Diagnoses are coded for each recorded hospital contact by primary diagnosis and when relevant secondary diagnosis. Surgical procedures were coded according to the Danish classification of surgical procedures from 1977-1995. Since then, surgical procedures have been coded according to a Danish version of the NOMESCO (Nordic Medico-Statistical Committee) Classification of Surgical Procedures (Sørensen et al. 2009).

Data in the DNPR are uploaded continuously and are received from PAS (the Patient Administrative Systems), a system in which all regions in Denmark collect and store information on the activities of hospitals in order to handle resource management. The Danish regions are required by law to submit standardised data at least monthly. The Danish Health Data Authority administers the DNPR and is responsible for maintenance and further development of the registry and performs routine checks of data prior to upload (e.g. missing codes, incorrect digits, errors in CPR numbers, inconsistencies between diagnosis and gender). It should be noted that registration of care provided by the private sector is mandatory, regardless of whether the referring hospital is public or private. However, reporting from private hospitals and clinics is generally considered incomplete (Schmidt et al. 2015).

Data quality: validity and completeness

Although the DNPR is considered to be a generally sound data source, both the content and definitions of single variables have changed over time, and for example changes in the organisation and provision of health services may affect both the type and the completeness of registrations. Basic information such as age, gender etc. has been included since the start of the registry, however, over time changes have been made in variables and classifications, and that has to be taken into account when using the data (Sørensen et al. 2009).

The validity of the data has been investigated from several perspectives including positive predictive value and intendent construct (Schmidt et al. 2015). The DNPR is a valuable tool for research however, since the validity of the data varies between variables, careful consideration should be made when using the registry.

Data quality in terms of completeness can be interpreted as the proportion of true cases of a disease that is correctly captured by the registry. Since no complete reference source exists, it is difficult to

estimate the overall completeness of registry data relative to the general population (Schmidt et al. 2015).

A systematic review has been performed of validated variables, based on several studies examining the data quality of individual variables (Schmidt et al. 2015). Large variation in data validity was found and underlines the need to validate diagnoses and treatments before using the DNPR data for research.

Accessibility of data, methodology and data linkage possibilities

Guidelines for the release of data from the DNPR have been established by the Danish Health Data Authority (previously part of the Danish Health and Medicines Authority). The Danish Act on Processing of Personal Data provides the legal basis for private and public institutions' collection of personally identifiable health data for research purposes. The act protects against the abuse of data, thus balancing the privacy rights of individuals against society's need for quality research. To access data from the DNPR (and other Danish national registries), researchers have to apply to Research Service (Danish, Forskerservice) (Schmidt et al. 2015; Danish Data Protection Agency 2017; The Danish Health Data Authority (webpage)).

As a research tool, the DNPR can potentially provide data for use in several study designs: cohort, case-control, cross sectional and ecological studies. Patient cohorts of interest may be identified with their medical history and outcomes. The DNPR may provide data on diseases, treatments and diagnostic examinations. Furthermore, DNPR allows for identification of disease occurrence in the general population (Schmidt et al. 2015).

There are several methodological issues that need to be considered when performing studies with data from the DNPR (Schmidt et al. 2015). Due to missing and incorrect data, selection bias and misclassification problems should be carefully evaluated. As with any observational study, there is always the possibility that unmeasured variables can affect the results. However, incomplete registration of some diagnoses and missing data on other characteristics may leave substantial residual and unmeasured confounding. As data in the DNPR have changed over time, a number of methodological problems particularly relating to disease incidence must be considered (Schmidt et al. 2015).

There is a huge potential for linking records to other Danish data sources using the CPR number as mentioned above. In this context, the Danish National Prescription Registry (Pottegård et al. 2015) is considered to be of central importance to epidemiological and pharmacoepidemiological research.

Over the years, a large number of studies and publications based on data from the DNPR, including linkage with other data sources, have been published. In recent years, applications have also involved major multinational networks and initiatives in which individual-level data of the DNPR have been integrated with data from other data sources, e.g. from other Nordic countries, as described previously.

Conclusion

The DNPR is a highly valuable tool and source for epidemiological research, providing longitudinal registration of diagnoses, treatments and examinations. The use of the civil registration number as identifier allows linking this registry to other data sources. Furthermore, since healthcare in Denmark is highly subsided by the government, several social classes are represented in the data. Records collect during the complete lifetime of the patient from birth to death. Thus, the DNPR is a unique source of

information for big data research. However, varying completeness and validity of the individual variables underline the need for validation of its clinical data before using the registry for research (Schmidt et al. 2015).

4.2.14.3. Appendix I.3.: European Society for Blood and Marrow Transplantation (EBMT)

Field: Blood and Marrow Diseases
Disease: Acute leukaemia, Amyloidosis, Bone marrow failure syndrome, Chronic myeloid leukaemia, Chronic lymphocytic leukaemia, Haemoglobinopathy, Inherited disorders, Juvenile idiopathic arthritis, Lymphoma, Multiple sclerosis, Myelodysplastic syndrome or md/mp neoplasm or secondary acute leukaemia, Myeloproliferative neoplasm, Plasma cell disorders including multiple myeloma, Solid tumour, Systemic lupus erythematosus, Systemic sclerosis.
Based in: The Netherlands
Website: www.ebmt.org

Background

The European Society for Blood and Marrow Transplantation (EBMT) is a non-profit organisation that was established in 1974 in order to allow scientists and physicians involved in clinical bone marrow transplantation to share their experience and develop co-operative studies. The EBMT registry is devoted to the promotion of all aspects associated with the transplantation of haematopoietic stem cells from all donor sources and donor types, including basic and clinical research, education, standardisation, quality control, and accreditation for transplant procedures.

The EBMT registry is the single biggest data source of its kind in Europe. The registry collects data in more than 500 centres and from more than 50 countries. It receives 30,000 new haematopoietic stem cell transplantation (HSCT) registrations per year and currently contains more than 500,000 HSCT procedures (https://www.youtube.com/watch?v=ecEqAkXqiu8).

The interests of the registry users are:

1) quality control of daily clinical care;

2) science and education;

3) market surveillance.

Dr. Jürgen Kuball (treasurer of the EBMT) recently gave a presentation on the Patient registries workshop of the EMA on 28 October 2016 (a video recording of the presentation is available via this link: http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/events/2016/08/event_detail_001315.jsp&mid=WC0b01ac058004d5c3)

In his presentation, Dr. Kuball explained that the data in the registry could be used to investigate the impact of a specific authorised medicine. The EBMT registry has been approached by several companies, owing partly to the efforts of the EMA, following which "The EBMT Non-Interventional Study Model" was developed.

The criteria for this model are:

- Non-interventional: Mild patient selection.

  Does not interfere with standard of care and the wide commercial use of the drug.

- Prospective:

  Collection of future events: for fixed time + additional follow up time.

- Study:

1) Generation forms and data bases per compound e.g. CAR-T cells, gene therapies of inherited disorders, etc.
2) Active data collection through:
MED-A (+ selected MED-B items)
MED-C.
3) Adding a control group (e.g. to calculate the incidence and compare different interventions).

Whether EC approval is needed will depend on whether the non-interventional prospective study (NIS) is initiated by academia or a pharmaceutical company. Whereas academia-initiated studies require EC approval in certain countries, studies initiated by pharmaceutical companies require EC approval in all countries. EMA has requested several non-interventional prospective studies to collect information about outcomes of autologous transplant in lymphoma and myeloma. These studies are performed in collaboration between the EBMT, industry and academia. For example, in the period 2011-2015 Genzyme and the EBMT performed a study to investigate the off-label use of plerixafor (study number NCT01362985).

Characterisation

Data structure, provenance of data and updates

The data are structured. Designated forms and manuals are used to enter the data. These forms can be found on the website of the EBMT – https://www.ebmt.org/Contents/Data-Management/Registrystructure/MED-ABdatacollectionforms/Pages/MED-AB-data-collection-forms.aspx.

The data are entered and maintained in a central database with internet access. Each EBMT centre is represented in this database, and users from a centre can enter, view, modify, obtain reports and download their own data once the necessary permissions have been granted by the principal investigator of the centre. In addition, all EBMT member centres can obtain general overviews of the complete EBMT data. The database is run and accessed through a system called ProMISe (Project Manager Internet Server).

National registries operating in some countries are integrated in the EBMT data flow by mutual consent and use the same central database. A small number of national registries enter data for their centres if preferred.

The data collected are reflected in the Med-AB forms. Centres can submit data by requesting data entry access to ProMISe and perform their own data entry (unless it is entered by their national registries).

Data quality: validity and completeness

- Exact definitions (harmonisation with US in progress).

  All items are completely defined before being placed in the data collection forms.
  Same items in different collection forms must mean the same.
  A definitions group made up of representatives of Working Parties and Study offices are always at hand to answer queries.

- Education and training.

  Training sessions for data managers on the use of ProMISe.
  Educational sessions on clinical knowledge specifically aimed at data managers.

- Database with internal quality controls.

Over 4,000 triggers control the accuracy and internal consistency of what is entered in the database at the point of entry.
Data quality reports can be run by users at any point to check for missing or unusual data.
Periodic queries on missing/incorrect data and follow up requests.

- Continuous support by the registry office.
  Helpdesk.

With the validation process the data heterogeneity is reduced to a minimum.

Accessibility of data, methodology and data linkage possibilities

The EBMT registry is the single biggest data source of its kind in Europe. The registry collects data in more than 500 centres and from more than 50 countries. It receives 30,000 new HSCT registrations per year and currently contains more than 500,000 HSCT procedures.

It is possible to obtain anonymous patient-level data, but most data requests require aggregated data.

Data can be accessed in collaboration with EBMT.

Regarding the accessibility of the centres, centre users can run columnar reports on their own data filtering the output by data items such as year of the HSCT, type of donor, diagnosis, etc. They can also run reports on aggregated data in the form of frequency tables or cross-tabulations. Centres that are members of the EBMT can also run reports on aggregated data from the whole database.

Any user with data entry access automatically has access to data retrieval.

If the user does not have data entry access, personal, non-transferable usernames and passwords can be requested for data retrieval only. As for data entry access, the request for data retrieval must be made to the Registry Office by faxing the form 'ProMIse personal password request – data download'. All individuals must be authorised by the Principal Investigator (PI) of the centre where they work. The name of the PI will be checked against the EBMT membership list. Data download access gives the user access to reporting and downloading the centre's data.

Unfortunately, there is not yet a unique identifier for an individual.

References (www.embt.org; Presentation on EBTM Registry at the Patient registries workshop at EMA 2016).

4.2.14.4. Appendix I.4.: European Cystic Fibrosis Society (ECFS)

Field: Rare diseases
Disease: Cystic fibrosis
Based in: Denmark
Website: https://www.ecfs.eu/

Background
The European Cystic Fibrosis Society (ECFS) is an international community of scientific and clinical professionals committed to improving survival and quality of life for people with cystic fibrosis (CF) by promoting high quality research, education and care.
The ECFS came into existence together with a new constitution, at the Annual General Meeting held during the 21st European CF conference in Davos, June 1997.

Purpose of the registry

To measure, survey and compare aspects of cystic fibrosis and its treatment in the participating countries, thereby encouraging new standards of dealing with the disease.
To provide data for epidemiological research.
To identify special patient groups suitable for multi-centre trials.

Deliverables

Continuing to create a network of European and International CF specialists including Allied Health Professionals to promote and stimulate the exchange of information about CF.
Holding annual conferences where specialists can meet and discuss all issues linked with CF. These conferences encourage the submission of research in the field to be presented in both oral and poster format.
Promoting young researchers.
Developing standardised European documentation for CF care.
Promoting the establishment of specialist Working Groups.
Publishing a Journal of CF (JCF) with six issues a year with supplements.

Introduction of the disease registry

The European Cystic Fibrosis Society Patient Registry was founded in 2004 and was based on the entry of defined demographic and clinical data.
Only patients who fulfil the diagnostic criteria below should be included in the registry:
a. Two sweat tests >60 mmol/L chloride.
b. One sweat test >60 mmol/L chloride AND DNA Analysis/Genotyping – two identified disease causing CF mutations.
If the sweat value is less than or equal to 60 mmol/L, then at least 2 of these should be fulfilled:
a. DNA Analysis/Genotyping – two identified diseases causing CF mutations.
b. Transepithelial (Nasal) Potential Difference – study consistent with a diagnosis of CF.
c. Clinical Presentation – typical features of CF.
Diagnosis reversal: if the patient's CF diagnosis reversed during the year, identify the reason from the options listed:
i. DNA Analysis – unable to identify two disease causing CF mutations.

ii. Transepithelial (Nasal) Potential Difference – study not consistent with a diagnosis of CF.
iii. Repeat normal sweat testing – confirm with clinical team.

Publications

Five articles were published or accepted for publication, and six abstracts were accepted in the period 2013 to 2015.
The ECFSPR data were handled in accordance with the ECFSPR guidelines.

Characterisation

Data structure, provenance of data and updates

Currently, the Registry includes demographic and clinical data of 38,985 consenting CF patients submitted by centres and national registries in the following 27 countries: Austria (9 individual centres), Belgium, Czech Republic, Denmark, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Latvia, Lithuania, Republic of Macedonia, Republic of Moldova, the Netherlands, Portugal (6 individual centres), Romania, Russian Federation, Serbia, Slovakia, Slovenia (2 individual centres), Spain (15 individual centres), Sweden, Switzerland (12 individual centres), Ukraine, and the United Kingdom. Altogether 27 applications requesting to use ECFSPR data have been submitted in the past three years (2013: 11, 2014: 7, 2015: 9).

Authorised users connect to the secure website and either input data by completing a web-based form, or they upload a data file in a compatible format e.g. Excel or XML. Only anonymised patient data, i.e. no patient or centre names, are sent to the ECFSPR database. Pre-agreed variable definitions, parameters and coding are used to allow reliable statistical analysis and reporting at centre, country and European level.

The ECFSPR collects the data through two methods: the use of spreadsheets converted to XML files and the use of a specific data-entry programme. If there is already an established registry/database in the patient's country/centre, it is possible to send the database without re-entering data by uploading the spreadsheet as an XML file using specifications that ECFSPR will provide, containing only the information that is compatible with the variables collected by the registry. The complete list of variables collected, and their coding are downloadable from the section "what are the variables collected" (https://www.ecfs.eu/sites/default/files/documents/Registry/Guidelines/VariablesDefintions_3_14.pdf) . The registry collects demographic and clinical data on all CF patients meeting the ECFS Patient Registry criteria.

Software/hardware

The ECFS Patient Registry uses a data-collection platform called ECFSTracker; an open source, multipurpose and multinational software program, custom-designed for the collection of cystic fibrosis patient data.
The data are stored at the University of Milan, Sezione di Statistica Medica e Biometria "G.A.Maccacaro". A server is located in secure premises, where access is allowed to authorised personnel only. Data storage conforms to Danish, Italian and EU data protection legislation and is approved by the Danish Data Protection Agency.

Data code

In order to ensure confidentiality of data, any user wishing to use the data-entry software must be assigned a code before data can be entered. For further information about code allocation, contact Hanne Vebert Olesen (hanne.olesen@ecfregistry.eu).

Data quality: validity and completeness

The quality of the data is guaranteed on several levels:
Level 1: At input level, in-built, automatic, controls and validation rules are applied by ECFSTracker. The software will block input or flag something that may be wrong, e.g. out of range values.
Level 2: Before data are transmitted to the ECFSPR, another series of controls is automatically applied by the software, and users have the possibility of modifying data.
Level 3: The final data checks are carried out by the ECFSPR Statistician.
National registries need to perform checks on their own data, as defined in a consensus document, before uploading their national data set to ECFSTracker. During upload and before transmission, the software will apply in-built checks and offer the opportunity to make corrections.

Accessibility of data, methodology and data linkage possibilities

Direct access to the data is allowed only to the CF centres. Only the doctor can enter and modify the patient's data. The biostatisticians in charge of data management and data analysis can see the database, but cannot modify patient data and cannot identify patients, because the patient's identity is protected by the unique code known only to the patient's centre. In case the biostatisticians need clarification on some of the data, they can send a query to the help desk (the only one able to link the centre code to the centre name). The help desk will then contact the centre. If data are entered directly from a single centre, the centres of a country can appoint a national coordinator who will then get access to anonymous data from all the centres in that country. This way the country will also have a national registry.
Researchers can apply for data for specific projects. The requests are reviewed by a scientific committee (which includes a representative appointed by the Cystic Fibrosis Europe patient organisation) to ensure that data are used according to the legislation and the aims of the registry as stated in the guidelines. If the permission is granted, the data will be analysed in cooperation between the researchers and the biostatisticians.
A data application form is available on the ECFS website (https://www.ecfs.eu/projects/ecfs-patient-registry/data-request-application).
All applications will be reviewed by the ECFS Scientific Committee. Applications from the pharmaceutical industry will also be reviewed by the ECFS Clinical Trials Network. Based on the recommendation of the ECFS Scientific Committee, the ECFS Steering Group (composed of national representatives of the countries that contribute data to the ECFSPR) will make a decision on the approval of the data request.
Applications from non-European countries must ensure an adequate level of protection (Directive 95/46/EC (TBC) chapter IV, article 25.1). Any application that involves anything other than aggregated data must be approved by the Danish Data Protection agency before release of data.
Applicants will be asked to sign an agreement in which they declare that the data will be used for the sole purpose indicated in the application and will not be kept for longer than necessary for the purpose(s) applied for. Reports, based on the data in the system for the centre, can be generated in real time. Graphs and tables can be downloaded and printed or visualised online at patient and centre level.

Data sharing: https://www.ecfs.eu/ctn

References

www.ecfs.eu; Excel spreadsheet: Gross list of registries (prepared by the EMA): Inventory of Registries – draft.xlsx

4.2.14.5. Appendix I.5.: European Registry for Multiple Sclerosis (EUReMS)

Field: Neurological diseases
Disease: Multiple sclerosis
Based in: Belgium

Website: http://www.eurems.eu/

Background

EUReMS is an EMSP project (2011-2014) on multiple sclerosis data collection, analysis and dissemination. It is focused on key concepts such as epidemiology, long-term therapy outcome, healthcare and quality of life of people with multiple sclerosis. The EUReMS project is co-funded by the European Commission under the Health Programme.

Purpose of the registry

The European Registry for Multiple Sclerosis, run from 2011-2014 by a consortium of academic institutions and NGOs, addresses the lack of data at EU and national level on treatment and care for people with multiple sclerosis (MS).
The EUReMS project is an initiative of the European Multiple Sclerosis Platform (EMSP) which represents those living with MS in Europe and has a network of 39 member societies in 34 European countries. As part of the EMSP's vision of a world without MS, the platform aims to improve quality of life as well as access to treatment, care and employment.
EUReMS' ultimate aim is to provide a comprehensive resource of collected data for research and practice for all European countries, including those that do not currently have their own. The aim for the post-2014 period is to use the newly created data infrastructure in collaboration with existing and emerging registries. This will eventually lead to a pan-European data pool to better assess the situation of people with MS.

Deliverables

EUReMS has identified and pooled MS-related data from different registries – hospitals, MS societies and research centres around Europe – and has created a cross-border partnership for its safe and effective storage, analysis, interpretation and dissemination. EUReMS data enables analysis of:

- costs and resources,
- age and gender-specific trends,
- disease-modifying drugs and their impact.

Introduction of the disease registry

EUReMS identified 20 MS registries across Europe; 12 of them started pooling their data in accordance with an agreed protocol to harmonise heterogeneous MS information.
The inclusion of the patients' perspective adds significant value to the project.
The first data pooling process was completed in August 2014 and formed the basis for four test studies addressing EUReMS' objectives:
EPI-1-d Study: Estimating Prevalence and Incidence of MS in Europe from EUReMS data collection, coordinated by Prof. M Pugliatti;
EPI-1-s Study: Comparison of the effect of the month of birth across Europe, coordinated by D Ellenberger and Prof. M Pugliatti;

DMD-1 Study: Comparison of access and effectiveness of DMD treatment for people with MS across Europe, coordinated by Prof. J Hillert;
PRO-1 Study: Assessment of people with MS' quality of life, the burden of disease and influence of employment from the patient's perspective across European countries, coordinated by Prof. P Flachenecker.

Publications

Large number of publications
http://www.emsp.org/resources/publications/

Characterisation

Data structure, provenance of data and updates

Until 2014, the EUReMS has identified 20 MS registries across Europe; 13 of them signed data sharing agreements. Data was collected from the following MS registries:
MS registry of Croatia, IMPULS MS Registry (Czech Republic), the Danish MS Registry, Tampere University Hospital Registry, Multiple Sklerose Registry der DMSG (Germany), Italian MS Database Network, MS registry of Liguria and Tuscany, Norwegian MS Registry and Bio bank, Polish MS registry, MS Registry of Serbia, Catalonian MS Registry, Svenska Multipel Skleros registret, UK MS Registry.

On the basis of the data shared by the mentioned registries, four studies were produced by the leading scientists involved in the project. Data was collected by questionnaires, telephone interviews with the registry leaders and on-site visits. 40 MS registries received the questionnaire; 23 of these registries completed it, and in 18 cases more detailed interviews were carried out to collect more details.

A core data set with 14 items, including date of birth, age at diagnosis, treatment received, quality of life and employment status has been established.
http://eurems.eu/attachments/article/93/EUReMS%20Data%20Mask_August2014.pdf

Software/hardware

Data gathering is managed by and stored at the University Medical Center Göttingen, Germany (UMG-GOE). Software tools for processing EUReMS data have been developed using the secuTrial database system at the UMG-GOE. This system also holds standard operating procedures (SOPs) detailing how the database is used and managed. The EUReMS database has been fully operational since May 2013 and follows national regulations and UMG-GOE policy.

Data code

EUReMS studies exclusively work with anonymised data.

Data quality: validity and completeness

In terms of data quality, especially comparability and integrity, a data handling routine has been implemented using an open source ETL (extract transform load) tool ("Talend Open Studio") to process the large amounts of heterogeneous raw data.

As a first step in harmonising datasets of different registries, a basic EUReMS data structure was defined for each of the four project studies, considering all information required to answer the research questions. Through the data handling process, the data exports are going to be converted into the prior defined study data structure to facilitate comparability and data analyses across the various registries participating in one study. In regard to quality assurance, the data handling process has been validated before providing data for analyses.
The data handling process consists of five steps: reading/splitting, cleaning, mapping and creating study datasets. During the first step, data is read and split into variables that are going to be used within the study datasets. The heterogeneity of the data is again noticeable in the data types of the source files, ranging from csv or Excel to Access Database. During the cleaning step, data is checked for incorrect or missing values and are, as a way of ensuring traceability, saved in specific reject files. In the mapping step, registry specific variables are mapped to the defined EUReMS denotations. By that, the heterogeneous data are harmonised, disabling misinterpretation of registry-specific variables, often in national language or unfamiliar abbreviations. The data is merged into study datasets that are uniform in appearance for each study and are provided to the statistical department for analyses in order to gain insight on disease related questions.

Accessibility of data, methodology and data linkage possiblilties

All data providers retain full ownership of contributed data, including the right to withdraw it. EUReMS holds ownership of compiled data. The EUReMS Board decides the rules for access and usage for each particular study.

Data sharing

Researchers and policymakers who wish to participate in the EUReMS studies' platform can apply to the EMSP Secretariat in Brussels, on the basis of access agreements and regulations developed by the EUReMS Steering Committee.

References (www.eurems.eu/; Excel spreadsheet: Gross list of registries (prepared by the EMA): Inventory of Registries – draft.xlsx)

4.2.14.6. Appendix I.6.: British Society for Rheumatology Biologics Register (BSRBR)

Field: Musculoskeletal conditions
Disease: Osteoarthritis, Rheumatoid arthritis, Spondyloarthropathies, Juvenile Idiopathic Arthritis, Crystal arthropathies, Septic arthritis, Lupus, Sjögren's syndrome, Scleroderma (systemic sclerosis), Polymyositis, Dermatomyositis, Polymyalgia rheumatica, Mixed connective tissue disease, Polychondritis, Sarcoidosis, Vasculitis
Based in: United Kingdom
Website: http://www.rheumatology.org.uk/

Background

The British Society for Rheumatology Biologics Register-Rheumatoid Arthritis (BSRBR-RA) is a national prospective cohort study that was established in 2001. It is a professional, multi-disciplinary, clinically led society representing healthcare professional members in rheumatology and musculoskeletal services in the UK and across the globe. Their members, following the integration with British Health Professionals in Rheumatology, include consultant rheumatologists, trainees, GPs and allied health professionals (AHPs). The British Society for Rheumatology (BSR) has a regionalised structure, with elected members across the UK, who work with commissioners to inform service design.

Purpose of the registry

BSRBR RA Register: The register monitors the long-term risks of serious adverse events over and above those that might be expected in patients treated with conventional therapy. Although the primary aim of the BSRBR is patient safety, a comprehensive range of data is collected.

BSRBR Ankylosing Spondylitis Register: The register recruits patients with ankylosing spondylitis who are being prescribed adalimumab, etanercept or certolizumab pegol as well as a control group of patients who have not been prescribed biologics; monitors the long-term safety of the treatments and increases understanding of their effects. It will include studying their efficacy, cost-efficacy, toxicity, adherence to guidance, and information on the effects of discontinuation of treatment or switching agents.

Deliverables

To improve awareness and understanding of arthritis and musculoskeletal conditions and highlight the role played by the multi-disciplinary rheumatology team in delivering high quality care to patients.
To make it easier for patients, clinicians and policy makers around the world to obtain the depth and quality of information they need on rheumatology.
To develop and promote standards of excellence to transform patient care and improve outcomes.

Introduction of the disease registry

The **BSRBR Rheumatoid Arthritis Register:** The RA register was set up in October 2001 and is the largest prospective register of rheumatology patients receiving anti-TNFα therapy in the world. It tracks the progress of patients with severe rheumatoid arthritis (RA), who are receiving biologic agents (adalimumab, anakinra, benepali, certolizumab pegol, etanercept, infliximab, rituximab and tocilizumab currently), monitoring the safety and effectiveness of these treatments over the long term.

The **BSRBR Ankylosing Spondylitis Register:** The AS register was launched in October 2012. The register recruits patients with ankylosing spondylitis who are being prescribed adalimumab or

etanercept, as well as a control group of patients who have not been prescribed biologics, in order to detect any long term or rare side effects.

Both registers are funded by the pharmaceutical companies which distribute the biologic therapies in the UK.

Publications

Over 40 publications have been submitted.

Characterisation

Data structure, provenance of data and updates

The **BSRBR-RA** is the largest prospective register of rheumatology patients receiving anti-TNFα therapy in the world. It currently has over 20,000 patients registered. Both patients and rheumatology health professionals complete BSRBR questionnaires on a six-monthly, then annual, basis. The register is supported by a team of 15 staff members at the Arthritis Research UK Epidemiology Unit at the University of Manchester (UoM).
BSRBR AS Register: Until May 2016, more than 1,300 people with AS have joined the register at the request of their rheumatology department.

For RA, both patients and rheumatology healthcare professionals submit data on a six-monthly basis, and the study is also linked to other national NHS databases (for instance, the UK cancer and death registries).

Consultant baseline questionnaire v11

The patients are contacted and asked to provide additional data on smoking habits and occupational history and are asked to complete a Health Assessment Questionnaire (HAQ) and the quality of life instrument Medical Outcome Survey Short Form 36 (SF-36).
All patients are followed for five years, including patients who stopped therapy or who switched to another biological agent. Three approaches to follow up are used:
Every six months the rheumatologist is surveyed asking for details of any changes in therapy, current disease activity (DAS 28) and specifically the development of any adverse event. Specific questions are asked about certain key events.
The patients are surveyed every six months (for three years) and complete a diary asking about any new diagnoses or significant comorbidities. All such reports from the patient or physician are followed by a request for more clinical information from the patient records. The focus is on serious adverse events defined particularly as leading to hospitalisation. All such events are recorded whether or not the physician attributes the event to the therapy.
All patients are flagged with the UK Office for National Statistics who then notify the register of: any death, with a copy of the medical information from the death certificate and any cancer.

Software/hardware
The BSRBR-RA is sponsored by the University of Manchester where the study is hosted. The BSRBR-AS is sponsored by the University of Aberdeen.
Software: Stata V.10 software (StataCorp, College Station, Texas, USA)

Data code

All adverse events are coded by the MedDRA scheme and reported to the sponsoring companies within 24 hours of receipt as well as in the form of six-monthly Periodic Safety Update Reports (PSUR).

Data quality: validity and completeness

Unknown

Accessibility of data, methodology and data linkage possibilities

BSBR encourages external parties to access and analyse the rich BSRBR-RA data set. However, there are some contractual limitations on how the data can be used. Requests to conduct research with the data follow a formal process (BSRBR Policy for third party data access). The length of time from request to approval and finally to supply of data can be as long as six months, depending on the complexity or size of the request.

BSRBR-AS

The Ankylosing Spondylitis dataset is limited. Before completing the request form, it is advised to discuss the requirements with Gary Macfarlane or Gareth Jones to ensure that AS register has the appropriate data. Requests to conduct research with the data follow a formal process as detailed above.

Data sharing

The register and its data are owned by the BSR, and the providers of the service have usual academic rights with regard to the data subject to the BSR's approval. This ensures an independence from the pharmaceutical companies that are funding the BSR to provide the register service.

References

www.rheumatology.org.uk/; Excel spreadsheet: Gross list of registries (prepared by the EMA): Inventory of Registries – draft.xlsx

4.2.15. References

The Anatomical Therapeutic Chemical (ATC) Classification System (WHO), website: http://www.who.int/medicines/regulation/medicines-safety/toolkit_atc/en/

Andersen, M. et al. "Implementing a Nordic Common Data Model for register-based pharmacoepidemiological research" https://www.ntnu.no/ojs/index.php/norepid/article/viewFile/1933/1907

Avorn J, "In Defense of Pharmacoepidemiology — Embracing the Yin and Yang of Drug Research", N Eng J Med, 2007; 357:2219-2221

Bouvy J.C., Blake K, Stattery J et al. "Registries in European post-marketing surveillance: a retrospective analysis of centrally approved products, 2005–2013", Pharmacoepidemiol Drug

Saf 2017; 26: 1442–1450.

Coebergh, J.W. et al. "EUROCOURSE recipe for cancer surveillance by visible population-based cancer RegisTrees in Europe: From roots to fruits", *Eur J Cancer*, 2015 Jun;51(9):1050-63.

Coebergh, J.W. et al. "EUROCOURSE lessons learned from and for population-based cancer registries in Europe and their programme owners: Improving performance by research programming for public health and clinical evaluation", *Eur J Cancer*, 2015 Jun;51(9):997-1017.

Cyanokit, EPAR, http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Summary_for_the_public/human/000806/WC500036362.pdf

Danish Data Protection Agency. *The Danish Act on Processing of Personal Data*. Available from: *http://www.datatilsynet.dk/english/the-danish-data-protection-agency/introduction-to-the-danish-data-protection-agency/*. Accessed May 25, 2017.

The Danish Health Data Authority. *Forskerservice*. Available from: *http://www.sundhedsdatastyrelsen.dk/da/forskerservice* (in Danish)

Danish Health and Medicine Authority. *(Evaluation of the Danish National patient registry 1990)*. Hospital Statistics II: 57; Copenhagen: 1993

Division of health planning at C.F. Møller on behalf of The Danish National Board of Health. Project concerning data quality in The Danish National Patient Registry (in Danish: Projekt vedrørende datakvalitet i Landspatientregistret). 2004.

Ehrenstein, V. et al. "Clinical epidemiology in the era of big data: new opportunities, familiar challenges", *Clinical Epidemiology* 2017: 9; 245-250

EMA website: http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000658.jsp

EMA: Initiative for patient registries – Strategy and pilot http://www.ema.europa.eu/docs/en_GB/document_library/Other/2015/10/WC500195576.pdf

EMA report: "Patient Registries Workshop, 28 October 2016 - Observations and recommendations arising from the workshop", http://www.ema.europa.eu/docs/en_GB/document_library/Report/2017/02/WC500221618.pdf

European Commission report: "Rare Diseases – A major unmet medical need" , 2017, https://ec.europa.eu/info/sites/info/files/rarediseases_p4p-report_2017.pdf

Excel-spreadsheet provided by EMA (Registry initiative): Inventory of Registries - draft.xlsx. Appendix.

Freemantle N, Walters K, Reynolds M, et al. "Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research", BMJ 2013;347:f6409

Furu, K. et al. "Selective serotonin reuptake inhibitors and venlafaxine in early pregnancy and risk of birth defects: population-based cohort study and sibling design", *BMJ*. 2015;350:h1798

Gini R. 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*. 2016;4(1):1189

Hastwell AJ, Baio G, Berlin JA, et al. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999–2014. BMJ Open 2016;6:e011666

International Classification of Diseases (ICD), website: http://www.who.int/classifications/icd/en/

Jürgensen, H.J. et al. "Registration of diagnosis in the Danish national registry of patients", Methods Inf MED, 1986; 25(3): 158-164

Kieler, H. et al. "Selective serotonin reuptake inhibitors during pregnancy and risk of persistent pulmonary hypertension in the newborn: population based cohort from the five Nordic countries", *BMJ*. 2012;344:d8012.

Kishnani, et al, "Recombinant human acid α-glucosidase: Major clinical benefits in infantile-onset Pompe disease", Neurology, 2007; 68;99-109.

Makady A. et al. "Policies for Use of Real-World Data in Health Technology Assessment (HTA): A Comparative Study of Six HTA Agencies", Value in Health, 2017; 20(4); 520-532.

Myozyme: EPAR Summary for the public. (http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Summary_for_the_public/human/000636/WC500032126.pdf)

Lynge, E. et al. "The Danish National Patient Register", *Scandinavian Journal of Public Health*, 2011; 39 (Suppl 7): 30-33

Pottegård, A. et al. "Data resource Profile: The Danish National Prescription Registry", *Intl J of Epidemiology*, 2015; 1-7

Presentation on EBMT Registry at the Patient Registries Workshop at EMA (http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/events/2016/08/event_detail_001315.jsp&mid=WC0b01ac058004d5c3).

Schmidt, M. et al. "The Danish National Patient Registry: a review of content, data quality, and research potential", *Clinical Epidemiology*, 2015; 7: 449-490

SNOWMED CT website: https://www.opencimi.org/tag/SNOWMED%20CT

Spigel D. R, The Value of Observational Cohort Studies for Cancer Drugs, Biotechnol Healthc., 2010; 7(2): 18–24.

Stephansson et al. "Selective serotonin reuptake inhibitors during pregnancy and risk of stillbirth and infant mortality", *JAMA*. 2013;309(1):48-54.

Soliris, EPAR http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Summary_for_the_public/human/000791/WC500054210.pdf

Suvarna V., Phase IV of Drug Development, Perspect Clin Res, 2010; 1(2): 57-60

Sørensen, HT et al. "Use of Medical Databases in Clinical Epidemiology", Department of Clinical Epidemiology, Aarhus university Hospital, 2009

Websites

http://www.encepp.eu/

http://www.caring-diabetes.eu/?q=content/general-introduction-0

https://www.ecfs.eu/

www.embt.org

http://www.eurems.eu/

www.eurocourse.org

www.iknl.nl

http://www.rheumatology.org.uk/

http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000658.jsp

Gross list of registries (prepared by the EMA):

https://www.ema.europa.eu/documents/report/inventory-registries_en.xlsx

4.3. *Drug consumption data (Sales and Prescription data)*

4.3.1. Background

Drug sales and prescription data provide information on the sales of medicines from manufacturers or wholesalers to pharmacies (community and hospital based) and retailers who are permitted to sell medicines, and the dispensing or selling of medicines from pharmacies to patients.

The IMI PROTECT project has already extensively reviewed the use, characteristics and availability of drug consumption data sources (sales and prescription) across the EU and is the source of much of the information contained in this report (Ferrer et al. 2011 & 2014).

The IMI PROTECT project used the following definition for drug consumption data sources (Ferrer et al, 2014):

- Sales: sales of medicines from wholesalers to community or hospital pharmacies and other retail outlets (sometimes termed "sell-in" data).
- Dispensed: medicines dispensed to patients in community pharmacies according to a prescription or obtained without a prescription (i.e. over the counter: OTC), (sometimes termed "sell-out" data).
- Prescribed: prescription medicines dispensed in community pharmacies and does not usually include OTC products. It may also include medicines prescribed and dispensed, but not reimbursed (e.g. oral contraceptives).
- Reimbursed: medicines reimbursed by health authorities or sickness funds prescribed by healthcare professionals, dispensed in a pharmacy and reimbursed by the healthcare provider. This does not include OTC medicines or non-reimbursed prescription-only medicines.

This overlaps slightly with the definition used in the electronic healthcare record data and registries section of this report where individual patient level prescribing and dispensing data are also captured.

4.3.2. Objectives

The purposes of this document are to identify relevant sources of European drug consumption data and describe the main characteristics of the data that impact on its use to support medicines regulation for conducting population-based observational studies.

4.3.3. Methods

The following approaches were used to obtain information describing drug consumption data sources:

- Review of the outputs of IMI PROTECT relevant to drug consumption databases.
- Literature search for additional relevant and selected data sources and the uses of them relevant to medicines regulation.

Based on this exploration, a general summary of the characteristics of relevant data sources is presented. Further, two specific examples of data sources in Europe, have been selected, for further, more in-depth characterization to illustrate the summary.

4.3.4. Data characterisation

4.3.4.1. Volume

IMI PROTECT conducted a comprehensive review of both commercial and non-commercial drug consumption data sources available across Europe (Ferrer et al. 2011 & 2014). They summarised information on 31 nationwide data sources of drug consumption data in 25 countries across Europe indicating high coverage of such data sources.

Drug consumption data sources usually provide aggregated data reflecting volume drug dispensed, sold or prescribed which limits their size. Patient level data are not usually available, although some databases offer further data at the individual patient level (e.g. Nordic countries and the Netherlands). The data are well structured.

There is a lesser volume of data available across the EU on drugs dispensed in hospital inpatient settings, which can be explained by the high heterogeneity in the management and distribution of medicines at a hospital level.

4.3.4.2. Veracity

As described, drug consumption data can be extracted at the point of wholesale, dispensing, prescription, or reimbursement. The data is well structured and generally highly processed before it is accessible for wider use.

Data sources vary with respect to coverage and may capture data on the entire population or be derived from a sample of the population; which is then, or can then be, projected up to a national estimate. The projection factor and sampling methodology will affect the validity and accuracy of drug sales estimates. Some smaller data sources may only be regional or cover certain settings or groups enrolled in specific health insurance plans.

Sales from wholesalers that include pharmacy stock movements and parallel trade may overestimate drug consumption. Conversely, reimbursement data may underestimate medicines consumption, as it does not include medicines available without a prescription or prescribed non-reimbursed medicines. The WHO recommends adjusting hospital drug consumption data by the level of clinical activity (e.g. adjusting for the number of occupied bed days and length of stay). However, this is not routinely carried out.

This report has not identified any standards or sets of rules on how or what data should be collected and thus the validity and accuracy of data may not be systematically assessed to the same degree as data from electronic healthcare records for example. Fully assessing the validity and accuracy of sales data are likely to be challenging and time-consuming given the size, coverage and difficulty in tracking medicines through the healthcare system. However, audits are likely to be conducted as part of the data collection process and should provide adequate confidence in the validity of the data.

4.3.4.3. Variability

As already discussed, there is variation across different data sources depending on when they extract data from the medicines distribution pathway.

There is also some heterogeneity in the definitions used for out- and inpatient drug consumption data. Some countries record the dispensation of hospital only medicines to outpatients as inpatient drug consumption, which may impact on the ability to compare data directly across countries. Drug consumption may also differ in some settings (e.g. institutionalised vs. community-dwelling elderly patients) and sales data may not be able to distinguish between such settings, which could be

important in assessing appropriate prescribing practices. The definition of healthcare setting needs to be considered carefully when conducting a comparison of drug consumption across multiple countries.

Data sources commonly use the Anatomical Therapeutic Chemical (ATC) classification system and the Defined Daily Dose (DDD) as developed by the WHO Collaborating Centre for Drug Statistics Methodology. The DDD allows for some comparisons to be made between databases and for the aggregation of data that differ in administration form and substance strength. There may be a delay in assigning DDDs to new drugs and therefore, for a study, DDDs may have to be estimated using data from RCTs or assigning the dose of the most frequently used strength. Data sources may express volume drugs sold or dispensed as the number of packs, number of tablets or number of millilitres. The DDD can then be used to estimate the number of patients exposed.

Some data sources use classification systems other than the ATC, such as the chapters of the British National Formulary (BNF) in the UK. BNF chapters have been approximately mapped to ATC classifications; however, each group may not strictly contain the same drugs.

4.3.4.4. Velocity

The rate of accumulation of data is contained and is largely dependent on the rate of increase in the population and to a lesser extent the number of licenced medicines.

4.3.5. Value

Data on drug sales and prescribing may be collected by ministries of health, government agencies, and healthcare and health insurance providers. Data may also be collected by the commercial sector. Usually the data are collected for purposes other than scientific research therefore there may be limitations with regards to the data captured and which impact its potential use. Data collected and collated by health ministries, government agencies and health service providers may be publicly available and free to use at no cost. Sales data are also available through the commercial sector, but these are usually subject to a cost to obtain access, which may preclude their use by researchers. Some datasets may require an application and approval process, which can delay access.

Data may be provided freely in excel spreadsheet or pdf files online. Other databases for which there is an associated cost may require software to be installed locally or accessed online to obtain access to, and to analyse, the data.

Aggregated sales data sets are typically relatively small and therefore easy to manage and analyse. Further simple analysis of the data may be required, however, to calculate defined daily doses and patient years of exposure for example.

Given most sales data are aggregated, they usually do not include information on variables considered as potential confounders. At most, some data sources may allow for stratification by age, gender and geographical region. Sales data are generally only used for descriptive rather analytical purposes.

Of note, IMI PROTECT also identified several public and private networks and working groups have been developed to promote the research on drug utilization through a collaborative international initiative with a variety of objectives. These include the European Drug Utilization Research Group (EuroDURG, https://www.pharmacoepi.org/eurodurg/).

4.3.6. Key Case Studies

The following specific data sources were selected to illustrate the two main types of drug consumption data.

IMS Health MIDAS – Customised Insights UK

Globally IMS Health MIDAS is available from over 90 countries and from over 140 medicines access channels. In IMS Multinational Integrated Analysis System (MIDAS), data are registered by drug and for all its application forms and it attempts to do this in a standardized way. Data collection is either sell-out, from pharmacy to consumer, or sell-in, from wholesale to pharmacy. In some countries, direct distribution from the manufacturer to pharmacy may also be captured.

Three sources of data are used to compile the data for the UK IMS MIDAS, as used by the MHRA, and a summary table is provided below outlining the main characteristics.

	Sector		
	Hospital	**Retail**	**Retail**
Coverage Summary			
Data Type	Consumption	Sell-Out	Sell-In
MIDAS Panel	UK Hospital	UK Retail	UK Sell In
Audit			
Audit Name	Hospital Pharmacy Audit (HPA)	UK Prescription Based Services (BPIX)	British Pharmaceutical Index (BPI)
Audit Type	Hospital to patient usage	Prescriptions dispensed to patients	Wholesaler and direct to pharmacy sales.
Frequency			
Data Availability	48 Quarters / 144 months	48 Quarters / 144 months	48 Quarters / Until M1213.
Frequency	Monthly	Monthly	Monthly
Data Coverage			
Primary Data Source	Hospitals	Pharmacies	Wholesalers
Secondary Data Source	-	Wholesalers (dispensed product)	Pharmacies, Manufacturers
IMS Sample of Channel	98%	78%	98%
Projection	Yes	Yes	Not projected
UK Audit Features			
Market Segmentation	Local and MIDAS	Local and MIDAS	Local and MIDAS
Molecule / Active substance	Yes	Yes	Yes
Therapy Class Classification	EphMRA ATC	EphMRA ATC	EphMRA ATC
Defined Daily Dose	No	No	No
Defined Days of therapy	No	No	No

The hospital consumption panel contains data from as early as 1991. The retail sell-in panel captures data from 1960, and dispensing data from dispensing doctors was included from 1982. This panel was no longer available after 2013. This was replaced by the retail sell-out panel which captured data from 2013 onwards.

The data is updated monthly but receiving the data from IMS can be by quarterly or monthly updates. The data is usually released 30-31 days after the end of that month's data collection.

The data is cleaned and projected to the UK population. Audit data collection and production is a local process i.e. UK IMS team. Core data collection items are all collected and reported via selected local audits. The core data is delivered to the MIDAS central data production and the data is linked and standardized to facilitate international comparison of markets using other IMS datasets.

The Hospital Pharmacy Audit (HPA) panel captures ~98.5% of NHS beds and provides a near complete capture of the NHS hospital sector. The hospital consumption audit does not cover private or military hospitals in the UK. Previously the retail sell-in panel captured ~98% of the retail channel which included the OTC products. As of 2013, the data unfortunately no longer captures sell-in data, therefore usage of OTC products bought in retail pharmacies is not available to the client via IMS MIDAS. In addition, medicines dispensed via local authority clinics (i.e. family planning clinics), internet pharmacies school and company medical centres, and home care channels (i.e. private nursing homes and hospices) or bought from supermarkets are not captured.

MIDAS can be customized to the needs of the client (in this case the UK MHRA) and a wide range of data elements are available for selection. Core data collection items are pack form, strength, size and volume, product name, the manufacturer, and number of packs sold/delivered through the channels (i.e. hospital and retail). The standard unit measure is number of packs and this is converted internally by IMS Health to the specific unit measures (kilograms, single units, counting units and international units as examples). The products are classified using the EphMRA Anatomical Classification of Pharmaceutical Products (ATC) and New form codes (NFC) classification system. At present, there is no DDD data element available, but it is planned to introduce this to the MIDAS database using the WHO ATC/DDD.

The dispensing data in MIDAS is not patient level therefore it is not possible to collect information on patient demographics, and specific treatment information such as dosage and duration of treatment. It is also not possible to link the data with other data sources. IMS Health offers several software packages analysing medical data and dispensing data, the medical channel is available via MIDAS or Prescribing Insights software which looks at medical, drug consumption and market data.

Clients are required to have a paid subscription for access to the data. National data extracted from this database could be shared with other NCAs based on agreed terms and conditions for data release of IMS data. The use of data in publications or communications are dependent upon agreed terms with IMS Health. Terms of use may be variable dependent on the client and the type of communications.

Danish National Health Registries – National Prescription Registry

Information on all prescriptions drugs sold in Denmark since 1994 has been recorded in the Register of Medicinal Products Statistics (RMPS). It contains individual-level information on dispensed prescriptions filled by Danish residents. Aggregated data on sales of OTC drugs and drugs sold for inpatient use is also captured and this is freely available online.

Following the legalisation of the use of individual-level data for research, prescription data has been made available to researchers since 2003 through Statistics Denmark and since 2014 the Danish Health Data Authority, Sundhedsdatastyrelsen. This sub-register is called the Danish National Prescription Registry (DNPR) and it contains anonymised individual-level data on all prescriptions dispensed at Danish community pharmacies. Data from 1994-2002 are not considered of sufficient quality to be used for research purposes and therefore not made available to researchers. It should be noted that, prescriptions dispensed for children under 16 years of age were classified under their mother's unique personal identification number (CPR) until 1996, and then the child's own CPR-number.

The DNPR contains individual-level data on prescriptions dispensed at community pharmacies for the entirety of Denmark. It also includes information on prescriptions dispensed to residents of long-term care institutions such as care homes. Tracking of the individual prescription history is based on the unique personal identification number – CPR number – which is assigned to all Danish residents at birth or upon immigration and is included in all national registers. DNPR captures data on dispensed prescriptions and not issued prescriptions. Dispensed prescription is a more reliable indicator for drug

usage than an issued prescription because a prescription issued by the physician may not be filled ("primary non-compliance").

There are 436 variables in the registry, and they can be divided into 4 main categories: drug user, prescriber, drug and pharmacy variables. The core variables in DNPR are the CPR-number, the dispensing date and the Nordic article number. The key variables are shown in Table 1.

The products are identified by the Nordic article number, which encodes information on trade name, formulation, strength, and package size. It is also linked to the WHO's Collaborating Centre for Drug Statistics, which makes it possible for ATC code, and the defined daily dose (DDD) to be identified. Unfortunately, DNPR does not capture drugs sold without a prescription i.e. OTC purchasing. If OTC drugs dispensed against a prescription for the treatment of a chronic disease, it would be recorded. In additional, prescriptions not dispensed at community pharmacies would not be captured in DNPR. This refers to drugs used in hospital admissions, drugs used by certain institutionalised patients and drugs supplied directly by hospitals or treatment centres. As the number of tablet or units per package and DDD is captured in DNPR it should be possible to standardize a unit measure for comparisons with other countries.

Table 1. Key variables in the DNPR (Pottegard et al. 2015).

Variable description	Variable name	Explanation
Patient details		
Personal identifier	CPR	Civil Personal Register (CPR) number, which encodes date of birth and gender and enables unambiguous linkage to other Danish registries
Dispensing details		
Date	EKSD	Date of completed sale/debit/dispensing
Packages	APK	Number of packets/units of the product dispensed
Product code	VNR	Product code of the product dispensed
Name*	PNAME	Product name
ATC*	ATC	WHO-defined Anatomical Therapeutical Chemical code
DDD*	VOLUME	Number of defined daily doses per package
Amount*	PACKSIZE	Number of tablets/units per package
Strength*	STRNUM	Numerical strength per tablet/unit
Form*	DOSFORM	Formulation of the drug
Other		
Prescriber	RECU	Identifier for the prescriber, e.g. a hospital or a general practice unit
Pharmacy	IBNR	Identifier for the dispensing pharmacy

*Within Statistics Denmark, these variables are included directly in the registry, whereas at the Danish Health Data Authorities they are obtained via linkage with the product code.

The data updates are dependent on which body the application for access for data is submitted to. The Danish Health Data Authority stores data from several health registries, including DNPR, and the data is made available with a delay of up to 2 months. The DNPR data stored within Statistics Denmark is updated twice annually and there is a delay of up to 9 months.

It is expected that the data has a high degree of completeness. All medicinal products are scanned using their bar code resulting in minimal data entry errors and there are financial incentives for pharmacies to completely register all purchases through the reimbursement scheme.

The original CPR-number is encrypted and replaced with a permanent identifier prior to the release of data to authorized researchers applying for the data. Linkage through the unique CPR-number enables researchers to link DNPR data with other Danish databases. The CPR-number has been assigned to all Danish residents since 1968 and the loss of follow-up is unlikely for permanent residents of Denmark. Any loss of follow up would be due to emigration and this can be traced.

An example of the benefit of linking data sources by the CPR number is a linked population-based database created for research on drug safety during pregnancy (Pedersen et al. 2016). This database

will use the CPR number to link data from the Danish National Registry of Patients, Danish Fetal Medicine Database, Danish Medical Birth Registry and the Danish National Health Service Prescription Database.

DNPR data is stored on servers within Statistics Denmark and Danish Health Data Authority. The data is only made available to users in anonymised form and cannot be accessed outside the two platforms. Researchers are authorised access for a specified period for specified subject-related purposes and they cannot transfer individual data to servers outside Statistics Denmark and the Danish Health Data Authority. All users must sign special confidentiality and non-disclose agreements in advance. Whilst the data is anonymized, further measures may be required if there is a risk of indirectly identifying single individuals, by removing or widening the definition of certain variables.

Access to the DNPR data is granted by application to the respective agency and a formal affiliation or collaboration with a Danish research institution is required. Only Danish research environments are granted authorisation. Foreign researchers can, however, get access to anonymised micro data through an affiliation to a Danish authorised environment. The DNPR data can only be used for analytical purposes and not under any circumstances used for administrative purposes. Overall approval regarding data protection is handled by both Statistics Denmark and the Research Services at the Danish Health Data Authority.

4.3.7. Conclusions

4.3.7.1. Potential use of drug consumption data throughout the Product Lifecycle

By its nature, drug consumption data is of most relevance for use within pharmacovigilance.

4.3.7.2. Impact of regulatory action and risk minimisation measures

The impact of regulatory action and implementation of risk minimisation measures, including product withdrawal (e.g. Hawton et al. 2009), medicine pack sizes (e.g. Hawkins et al. 2007) and communication of drug safety alerts (e.g. Herdeiro et al. 2016) have been studied using drug consumption data. Assessing the impact of regulatory action and risk minimisation measures is an integral part of the pharmacovigilance cycle. Drug consumption data can be used to provide rapid assessment of the effectiveness of such measures and help in the redirection of efforts should measures have been ineffective. However, these data will generally only provide information on changes in sales trends and electronic healthcare record data are likely to be required should any additional clinical data be needed.

4.3.7.3. Impact of reclassification of medicines

Several studies have studied the impact of changes to legal status of medicines following reclassification to OTC availability on sales and prescribing (e.g. Dhippayom & Walker. 2006, Walker & Hinchliffe. 2010, Du et al. 2014). Rapid assessment of any changes in drug use can be assessed and the impact on other drug classes, if any, can also be investigated.

4.3.7.4. Prescribing trends, utilisation and market uptake

Studies have used drug consumption data to assess trends in prescribing for multiple drug classes across multiple EU countries (e.g. Walley et al. 2005). Trends in consumption following the introduction of new formulations (e.g. Treceno C et al. 2012) and the marketing of biological medicines (e.g. Obradovic et al. 2009) have also been studied. The data obtained from these types of studies are of interest to help establish the extent of use of a medicine in the population and help prioritise

actions. These data are also of value to pharmaceutical companies for marketing purposes and for use in regulatory submissions when required.

4.3.7.5. Signal assessment of adverse drug reactions

Drug consumption data have been used to supplement spontaneous adverse drug reaction reports to assess the impact of regulatory measures (e.g. Motala et al, 2008) and changes in prescribing practice (e.g. Khong et al, 2012) on reporting rates and to put reports of adverse drugs reactions into context (e.g. Jonville-Bera et al, 2011) including through use of ecological or case-population study designs (e.g. Gulmez et al. 2013). These approaches may be of value, particularly if there is limited data available from other sources. These studies are ecological in design and will therefore suffer from bias but can generate hypotheses for studying potential associations in other databases.

4.3.7.6. Comparative drug expenditure

Spending on orphan drugs (e.g. Orofino et al, 2010) and generic drugs (e.g. Wouters et al, 2017) have been compared across EU countries as well the impact of economic policies on drug utilisation (e.g. Leopold et al, 2014). The application of sales data to these studies is of relevance to industry and for Health Technology Appraisal (HTA).

4.3.7.7. Disease surveillance

Drug consumption data have been used for disease surveillance purposes including the detection of infectious disease epidemics (e.g. Pivette et al, 2014) and the impact on disease rates (e.g. Viola et al, 2008). The use of drug consumption data for these purposes may be of use to assess the potential public health impact that medicines have on disease rates and be of interest to health ministries and public health agencies.

4.3.8. Regulatory challenges

Drug consumption data can be useful tools however, they present several challenges, which may require seeking data from additional sources, including linked electronic health records, and their range of potential uses in small. They have some value for regulators themselves but will not feature except on rare occasions within regulatory submissions aside from in relation to the sales or projected sales of the particular product of interest.

Studies utilising such data tend to be ecological in design and will therefore be subject to the ecological fallacy bias. The results are usually only descriptive and can only be used to generate hypotheses rather than test them. When combined with other data sources, including adverse reaction reports, drug consumption data can put such reports into context, help prioritise regulatory measures and assess the public health importance of potential risks.

Given these data are often aggregated and not patient level, there are likely to be limited or missing information on potential confounders and patient characteristics. There is limited, if any information on the indication and this will be compounded, particularly if a medicine has more than one licensed indication. There are likely to be limited capabilities for linking to other sources, including electronic healthcare records, which can limit their use for exploring indication etc.

There is also limited availability of individual patient-level data from hospital or specialist settings as in many countries there is less networking of databases in these environments.

Sales data available through commercial providers may come at a considerable cost, which can preclude their use in the event of budgetary constraints. The timeliness of data availability is another

factor that affects the value of sales data, particularly if up to date near real-time data are required for regulatory purposes. The sharing and inclusion of sales in external reports and press releases may require approval from data owners and custodians, which can cause delays in communicating regulatory messages to stakeholders.

4.3.9. Recommendations

Drug consumption data are in general comprehensively accumulated and processed. As highlighted, there are some limitations of the data currently available, notably the limited availability of individual patient-level prescribing data particularly from hospital in-patients; however, this is generally a result of the medicines access route and current IT capabilities within those settings. Initiatives to connect in-patient hospital data so that it can be included in prescribing databases are likely to be extremely difficult although this could be considered across smaller connected groups of hospitals and should be an area of high priority. This is also relevant when considering electronic healthcare record databases.

Comparison of data across different countries is also likely to be of interest and standardisation of the data across countries helps facilitate this. A good practice guide manuscript for the conduct of multi country drug utilisation studies, to be endorsed by the International Society of Pharmacoepidemiology (ISPE), is currently under development. Adoption of the ISO IDMP standards for the identification of unique products would likely be beneficial particularly if these standards were also taken up by electronic healthcare record databases as recommended above.

It is recommended that regulators have consistent and easy access to drug consumption data, meaning that an inventory of such sources should be maintained, and that there is expertise available to analyse it in-house if necessary, as it can be a useful resource. In particular, consideration should be made by NCAs as to how they can optimally use this data to routinely support signal assessment within pharmacovigilance by placing spontaneous adverse event reports into context and how it may be used to routinely monitor the actual or potential impacts of regulatory action with experiences of using it shared.

4.3.10. References

Ferrer P, Ballarin E, Sabate M, et al. on behalf of the PROTECT project. Drug consumption databases in Europe. Barcelona, August 2011. Available at http://www.imi-protect.eu/documents/DUinventory_2011_6_WORD97-2003.pdf [Accessed June 2017].

Ferrer P, Ballarin E, Sabate M, et al. Sources of European drug consumption data at country level. Int J Public Health 2014; 59(5): 877-87.

Dhippayom T, Walker R. Impact of reclassification of omeprazole on the prescribing and sales of ulcer healing drugs. Pharm World Sci 2006; 28(4): 194-8.

Du HC, John DN, Walker R. An investigation of prescription and over-the-counter supply of ophthalmic chloramphenicol in Wales in the 5 years following reclassification. Int J Pharm Pract 2014; 22(1): 20-7.

Sulmez SE, Larrey D, Pageaux G-P, et al. Transplantation for acute liver failure in patients exposed to NSAIDs or paracetamol (acetaminophen). Drug Saf 2013; 36(2): 135-44.

Hawkins LC, Edwards JN, Dargan PI. Impact of restricting pack sizes on paracetamol poisoning in the United Kingdom: a review of the literature. Drug Saf 2007; 30(6): 465-79.

Hawton K, Bergen H, Simkin S et al. Effect of withdrawal of co-proxamol on prescribing and death from drug poisoning in England and Wales: time series analysis. BMJ 2009; 338: b2270.

Herdeiro MT, Soares S, Silva T et al. Impact of rosiglitazone safety alerts on oral antidiabetic sales trends: a countywide study in Portugal. Fundam Clin Pharmacol 2016; 30(5) 440-9.

Jonville-Bera AP, Autret-Leca E. Adverse drug reactions of strontium ranelate (Protelos(®) in France. Presse Med 2011; 40(10): e453-62.

Khong TP, de Vries F, Goldenberg JS, et al. Potential impact of benzodiazepine use on the rat eof hip fractures in five large European countries and the United States. Calcif Tissue Int 2012; 91(1): 24-31.

Leopold C, Zhang F, Mantel-Teeuwisse AK. Impact of pharmaceutical policy interventions on utilization of antipsychotic medicines in Finland and Portugal in times of economic recession: interrupted time series analyses. Int J Equity Health 2014; 13: 53.

Motala D, Vargiu A, Leone R, et al. Influence of regulatory measures on the rate of spontaneous adverse drug reaction reporting in Italy. Drug Saf 2008; 31(7): 609-16.

Obradovic M, Mrhar A, Kos M. Market uptake of biologic and small-molecule—targeted oncology drugs in Europe. Clin Ther. 2009; 31(12): 2940-52.

Orofino J, Soto J, Casado MA et al. Global spending on orphan drugs in France, Germany, the UK, Italy and Spain during 2007. Appl Health Econ Health Policy 2010; 8(5): 301-15.

Pedersen LH, Petersen OB, Norgaard M, et al. Linkage between the Danish National Health Service Prescription Database, the Danish Fetal Medicine Database, and other Danish registries as a tool for the study of drug safety in pregnancy. Clinical Epidemiology. 2016; 8: 91-95.

Pivette M, Mueller JE, Crepey P, et al. Surveillance of gastrointestinal disease in France using drug sales data. Epidemics 2014; 1-8.

Pottegard A, Johannesdottir Schmidt SA, Wallach-Kildemoes H, et al. Data Resource Profile; The Danish national Prescription Registry. Int J Epidemiol. 2016.

Treceno C, Martin Arias LH, Sainz et al. Trends in the consumption of attention deficit hyperactivity disorder medications in Castilla y Leon (Spain): changes in the consumption pattern following the introduction of extended release methylphenidate. Pharmacoepidemiol Drug Saf. 2012; 21(4): 435-41.

Viola R, Benko R, Nagy G, Soos G. National trend of antidepressant consumption and its impact on suicide rate in Hungary. Pharmacoepidemiol Drug Saf 2008: 17(4): 401-5.

Walker R, Hinchcliffe A. Prescribing and sales of ophthalmic chloramphenicol following reclassification to over-the-counter availability. Int J Pharm Pract 2010; 18(5): 269-74.

Walley T, Folino-Gallo P, Stephens P et al. Trends in prescribing and utilization of statins and other lipid lowering drugs across Europe 1997-2003. Br J Clin Pharmacol 2005; 60(5): 543-51.

Wouters OJ, Kanavos PG. A comparison of generic drug prices in seven European countries: a methodological analysis. BMC Health Serv Res 2017; 17(1): 242.