Unmet Big Data needs in the health care system

IMI2 Project ID – DO->IT

Big Data for Better Outcomes, Policy Innovation and Healthcare System Transformation

WP1 – Programme Strategy and Coordination

Lead contributor	London School of Economics and Political Science (LSE) Maximilian Salcher; Caroline Carney; Huseyin Naci
Other contributors	Imperial College London (ICL) Katharine Smith; Søren Rud Kristnesen
	Swedish Institute for Health Economics (IHE) Peter Lindgren; Fredrik Moen; Niklas Wallin Bernhardsson
	Centre for Research in Healthcare Management (CERGAS) Helen Banks; Michela Meregaglia; Giovanni Fattore
	National Institute for Health and Care Excellence (NICE) Pall Jonsson; Heather Stegenga
	GlaxoSmithKline (GSK) Bart Barefoot; Marianne Cunnington
	Bayer (BAY) Matthias Gottwald
	Pfizer Carlos Nunes; Margaret M McDonald; Jeremy Hayter

Delivery date	31 Oct 2018

Table of Contents

Executive summary 3
List of acronyms 6
1 Introduction 7
1.1 Fit for purpose? Big data for regulatory and health technology assessment 7
1.2 Unmet big data needs for health care decision-making 8
2 Data for decision-making: needs, availability, and methods for analysis 9
2.1 Data needs and availability for regulatory and HTA decisions: a case study approach 10
2.1.1 Overview of identified data sources 12
2.1.2 Overview of case studies 12
2.1.3 Regulatory and HTA data needs summary 13
2.1.4 Limitations 14
3 Future research priorities 15
3.1.1 Putting big data to use: overcoming barriers and demonstrating value 15
3.1.2 Making patient-reported outcomes available in routine data sets 16
3.1.3 Increasing acceptability by demonstrating trade-offs 17
3.1.4 Real-world patients: more to be learned from big data 18
4 References 20
Appendix 24
Case study I Multiple sclerosis Potential value of big data for the approval of alemtuzumab 26
Case study II Inflammatory bowel disease Potential value of big data for the approval of infliximab 31
Case study III Lung cancer Potential value of big data for the approval of crizotinib 41
Case study IV Multimorbid patients Potential value of big data for the approval of medicines in chronic conditions 45
4.1 General (non-disease specific) data sources 60
4.2 Disease-specific data sources: multiple sclerosis 88
4.3 Disease-specific data sources: inflammatory bowel disease 95
4.4 Disease-specific data sources: lung cancer 99

Executive summary

The potential of 'big data' in health care has created excitement among researchers and health care decision-makers alike. Availability of more data from routine data sources allows insights into practice variation, patterns of diseases, safety and effectiveness of treatments in real-world situations. The IMI2 Big Data for Better Outcomes (BD4BO) initiative is a comprehensive research programme that aims to develop key enablers to support health care system transformation through the use of big data in a range of exemplary disease areas. The programme is based on the understanding that availability of better, more integrated data will allow more insightful assessment of health system needs and health care decision-makers to act on this, resulting in improved health outcomes and health care systems in Europe.

Big data can impact on various aspects of the health care system, including the work of medicinal product regulators and health technology assessment (HTA) bodies. These decision-makers require valid and robust information on whether new medicines work in the populations they are indicated for. While there appears to be general agreement about the potential of big data to support regulatory and HTA decision-making, empirical assessments of this claim are still largely missing. Establishing what is available and to what extent available data can meet requirements for decision-making can help identify unmet big data needs and inform future investments in research in this area to support a move towards a data-driven, evidence-based health care system. Once data are available, questions around their validity (quality of collected data) and how these data are analysed become important. Methods of statistical analysis of big health care data that allow causal inferences to be drawn are of particular interest.

A three-pronged approach was employed to identify potential unmet big data needs for health care decision-making and to inform future investments in research in this area to support a move towards a data-driven, evidence-based health care system.

- First, we evaluated what's available and what's missing in terms of data from routine European data sources. We conducted a scoping review of routine data sources in four exemplary disease areas (lung cancer, multiple sclerosis, inflammatory bowel disease, multimorbidities) in seven European countries (Finland, France, Hungary, Italy, Norway, Sweden, United Kingdom) and then assessed available data sources against data needs of health care decision-makers as evidenced by data used for previous decisions and data needs communicated in official documents (EMA European Public Assessment Reports, national HTA reports).
- Second, we asked the question what can we do with the data? Novel methods that are typically used in the econometric literature and have made inroads to becoming tools in comparative effectiveness research as well were reviewed. Under certain assumptions, these methods allow causal inference to be drawn from nonexperimental data to inform decision-making.
- Finally, we drew upon the experience and knowledge of national and international policy-makers, payers, HTA bodies, academics, and patient representatives in the form of members of the IMI2 DO→IT International Advisory Board. A workshop with this group was held where the role of big data for health care decision-making and its implications for future research were discussed.

A total of 164 generic and disease-specific data sources were identified in seven European countries. These comprised of various administrative data bases (such as social insurance data bases, claims data bases, prescription data), death registers (with cause of death), biobanks, surveys, and disease-specific registries for the selected disease areas. Potential contributions of evidence from routine data bases to fill gaps identified in regulatory and health technology assessment reports can be categorised into three main areas: safety

(often in relation to populations excluded from pivotal trials), efficacy/effectiveness (again, in relation to real-world populations), and understanding drug use patterns (such as treatment adherence and use of concomitant medication). The case studies also indicate that despite increasing excitement about its potential, observational evidence may not yet have found its place to address regulatory and health technology assessment uncertainties that goes beyond existing uses for these data sources (such as post-authorisation safety monitoring).

A review of the data and design considerations as well as strengths and limitations of econometric methods to assess causal associations in observational data showed that although randomised controlled trials (RCTs) continue to be perceived as an effective means of establishing cause-effect relationships, a range of alternative methods exists for evaluating non-experimental data. These methods can address limitations of RCTs such as the demand on resources of large size RCTs studies and long follow up times for certain outcomes which can be alleviated through complementary use of real-world data. Additionally, causal analysis of real-world data can address issues such as RCTs not being feasible due to small patient populations or ethical concerns regarding randomisation for promising treatments. Each of the econometric methods reviewed can, once certain assumptions are met, provide evidence of causal associations. Studies applying these methods can also benefit from improved external validity in particular in relation to populations commonly excluded from RCTs such as those with comorbidities.

Unmet big data needs identified in this exploratory study relate to the availability of relevant information as well as its analysis. The discussion of this work with the DO→IT IAB resulted in the following recommendations:

- *The future challenge of big data research is not a lack of data sets, but making sure that existing and prospectively collected data is put to use. Greater use requires that data is of high quality, can be easily identified and robustly linked, and that the value of doing so is understood by key stakeholders. There is a need for further research that enables readily identification of existing big data sources and assessment of their quality, sets minimum standards that data collectors can choose to comply with to allow robust linkage at the individual level, and clearly articulates the opportunity costs of not linking data. This should be complemented with clear procedures for obtaining consent for data usage.*

- *Although a vast number of data sources are available, an important data gap exists for patient-reported outcomes. Patient reported outcomes are becoming increasingly important for the assessment of new treatments but are typically not collected routinely. There is a need to strengthen research into disease-specific minimum sets of patient-reported outcomes that should be collected in routine practice.*

- *A major necessity for further uptake of big data and non-experimental methods in the regulatory process is cultural change. Research that demonstrates the advantages and limitations of non-experimental methods for causal inference, and identifies situations when non-experimental evidence can complement RCTs, can help build the cultural change necessary to increase the acceptability of non-experimental evidence by regulators and increase the incentives for companies and academics to share their data*

- *While there are challenges to including patients reflecting the real-world patient population in traditional clinical trials, pragmatic trials could be used to investigate effectiveness and safety using robust methodological standards in real-world populations as they make use of existing data collection infrastructure while retaining the randomisation process to control for confounding factors.*

List of acronyms

BD4BO - Big Data for Better Outcomes
DID - Difference-in-differences
EMA - European Medicines Agency
FDA - Food and Drug Administration
HAS - Haute Autorité de Santé
HTA - Health technology assessment
IBD - Inflammatory bowel disease
IMI2 - Innovative Medicines Initiative 2
IV - Instrumental variables
ITS - Interrupted time series
MR - Mendelian randomisation
MS - Multiple sclerosis
NICE - National Institute for Health and Care Excellence
PSM - Propensity score matching
RCT - Randomised controlled trial
RDD - Regression discontinuity design
RWD - Real-world data
RWE - Real-world evidence
TLV - Swedish Dental and Pharmaceutical Benefits Agency

1 Introduction

The potential of 'big data' in health care has created excitement among researchers and health care decision-makers alike. Availability of more data from routine data sources allows insights into practice variation, patterns of diseases, safety and effectiveness of treatments in real-world situations. Commentary pieces and reviews (e.g., Murdoch and Detsky, 2013; Roski, Bo-Linn and Andrews, 2014; Weber, Mandl and Kohane, 2014; Berger *et al.*, 2015; Salas-Vega, Haimann and Mossialos, 2015; Salcher, 2017) have highlighted opportunities and challenges arising from the 'big data revolution' (Groves *et al.*, 2013) and a variety of potential uses of big data have been suggested, including system-level (monitoring of healthcare service delivery, including identification of over-use and under-use), disease level (enhancing our understanding of the natural progression of a disease and the target population for new medicines; identifying the occurrence of co-morbidities through disease surveillance systems; enabling the detection of population-level effects; identifying high-value treatments), and trial/product level applications (efficient and sensitive recruitment of patients for clinical trials; real-time learning and monitoring; re-use of existing data; provide source of pragmatic, real-world, evidence on the effectiveness and cost-effectiveness of treatments). However, empirical assessments of the value of available data for these applications is still often missing and research efforts to demonstrate relevance and robustness of big data for health care decision-making is often fragmented.

The IMI2 Big Data for Better Outcomes (BD4BO) initiative is a comprehensive research programme that aims to develop key enablers to support health care system transformation through the use of big data in a range of exemplary disease areas[1]. The programme is based on the understanding that availability of better, more integrated data will allow more insightful assessment of health system needs and health care decision-makers to act on this, resulting in improved health outcomes and health care systems in Europe.

1.1 Fit for purpose? Big data for regulatory and health technology assessment

Having been ascribed transformative potential for the health care landscape (Groves *et al.*, 2013), big data can impact on various aspects of the health care system (see above), including the work of medicinal product regulators and health technology assessment (HTA) bodies. These decision-makers require valid and robust information on whether new medicines work in the populations they are indicated for. 'Big data' from routine data bases can provide valuable insights, as they reflect the real world better than traditional clinical trials, which are often conducted in restricted populations. Provided that high standards for data collection, study design and data analysis can be achieved, these data can be helpful for assessing real-world effectiveness and safety of treatments – key aspects to consider when making far-reaching decisions about the allocation of scarce resources in the health care system.

Various definitions of big data exist, all describing a similar concept: large data sets of diverse origin and format, that contain information requiring novel methods to be processed and analysed. The big data promise for regulatory and HTA decision-making largely builds on the opportunities provided by routine data sources that allow to study how medicines work in the real world or in situations where traditional clinical trials are challenging to conduct. The following definition developed for a study commissioned by the European Commission is useful in this context:
"Big Data in Health refers to large routinely or automatically collected datasets, which are electronically captured and stored. It is reusable in the sense of multipurpose data and comprises the fusion and connection of existing databases for the purpose of improving

[1] The field of big data research is rapidly developing BD4BO does not exist in a vacuum. Recent related project of particular relevance include the IMI funded GetReal and PROTECT projects.

health and health system performance. It does not refer to data collected for a specific study." (Habl *et al.*, 2016) Although this definition does not explicitly mention data sources such as wearables and genomics, we shall take it to also include these, in that they are automatically collected, and recognise that while individual data sets are not necessarily "big" data, they can fall within the definition via linkage of data sets that together form a "big" data set.

While there appears to be general agreement about the potential of big data to support regulatory and HTA decision-making, empirical assessments of this claim are still largely missing. Establishing what is available and to what extent available data can meet requirements for decision-making can help identify unmet big data needs and inform future investments in research in this area to support a move towards a data-driven, evidence-based health care system. Much data is already available in European health care systems, but often in silos, and it is unclear whether available routine data bases are suitable to build the basis for decision-making in the health care system by providing required information to conduct analyses of efficacy, comparative effectiveness and comparative cost-effectiveness. A clear potential for more efficient use of available data lies in the re-use of existing data, as evidenced by the findings of a recent review that showed that over a third of registries used for post-marketing surveillance purposes in Europe are drug registries, collecting data exclusively on patients receiving the drug under surveillance (Jonker *et al.*, 2017).

Data availability is a pre-requisite for data-driven health care system transformation. Once data are available, questions around their validity (quality of collected data) and how these data are analysed become important. Methods of statistical analysis of big health care data that allow causal inferences to be drawn are of particular interest. With increasing use of observational data for regulatory and health technology assessment, methods such as propensity scores, instrumental variables, regression discontinuity design and others emanating from the econometric literature could become increasingly important tools and their relevance to analyse big data for health care decision-making needs to be assessed.

1.2 Unmet big data needs for health care decision-making

In this report, considerations about unmet 'big data' needs are developed, based on the understanding that big data will take a prominent role in a data-driven, evidence-based health care system, and that the IMI2 BD4BO initiative can provide important input to support this development. Questions about the availability of data in European health care systems are raised, and case studies in selected disease areas provide some insights into where data gaps exist. Methods for analysing large data sets originally used in the econometric literature are reviewed for health care research. Finally, a set of recommendations to address unmet big data needs for health care decision-making is developed based on the reviews conducted for this report, and the input provided by the International Advisory Board of the IMI2 BD4BO Coordination and Support Action, DO→IT, comprising national and international policy-makers, payers, HTA bodies, academics, and patient representatives.

2 Data for decision-making: needs, availability, and methods for analysis

A three-pronged approach was employed to identify potential unmet big data needs for health care decision-making and to inform future investments in research in this area to support a move towards a data-driven, evidence-based health care system.

- First, we evaluated what's available and what's missing in terms of data from routine European data sources. We conducted a scoping review of routine data sources in four exemplary disease areas (lung cancer, multiple sclerosis, inflammatory bowel disease, multimorbidities) in seven European countries (Finland, France, Hungary, Italy, Norway, Sweden, United Kingdom) and then assessed available data sources against data needs of health care decision-makers as evidenced by data used for previous decisions and data needs communicated in official documents (EMA European Public Assessment Reports, national HTA reports).
- Second, we asked the question what can we do with the data? Novel methods that are used in the econometric literature and have made inroads to becoming tools in comparative effectiveness research as well were reviewed. These methods allow causal inference to be drawn to inform decision-making.
- Finally, we drew upon the experience and knowledge of national and international policy-makers, payers, HTA bodies, academics, and patient representatives in the form of members of the IMI2 DO→IT International Advisory Board. A workshop with this group was held where the role of big data for health care decision-making and its implications for future research were discussed.

The details of the case studies in four disease areas, identified data sources in seven European countries, and a full report on econometric methods for causal inference can be found in the **appendix** of this report.

Below, the methods for this work and key results are outlined. First, a **brief overview of the methods and summary of findings** from identification and mapping of data sources and needs of HTA bodies for our case studies are presented. Then, an overview of the methodological reasons that make causal inference methods for analysing big data sets is given, and **key methods from the econometric literature** with their advantages and limitations are presented. Finally, **key themes emanating from case studies, methodological work and the consultation with the DO→IT International Advisory Board** are discussed, and a series of **recommendations** for future research are developed.

2.1 Data needs and availability for regulatory and HTA decisions: a case study approach

Empirical assessments of the value of available data for regulatory and HTA decision-making are still largely missing to inform future investments in research in this area and to support a move towards a data-driven, evidence-based health care system. It is unclear whether available routine data bases in Europe are suitable to build the basis for decision-making in the health care system by providing required information to conduct analyses of efficacy, comparative effectiveness and comparative cost-effectiveness.

A team of researchers from the IMI2 BD4BO DO→IT consortium therefore set out to identify big data sets across Europe and assess their potential value for producing evidence as basis for far-reaching decision about making treatments available to patients and allocating funds. This exploratory study takes a case-study approach. Available data sets in seven European countries were reviewed, and the information they contain (coverage, outcomes, covariates) was assessed for its usefulness for informing decisions about new medicines in four selected, representative disease areas. Overall, our approach was three-fold: for each of the selected disease areas, we first identified available data sources. We then identified the data needs of regulators and HTA bodies by reviewing assessment reports for case study drugs in the four disease areas. As a last step, we compared available data with identified data needs. In the box below, an overview of the methods used is given.

A set of four exemplary disease areas were selected as case studies, including: multiple sclerosis, inflammatory bowel disease, lung cancer, and multimorbid patients. These disease areas were selected because they were deemed amenable to improvement in research, medicines development and achievement of better outcomes through the use of routine data by members of the BD4BO DO→IT consortium, including HTA bodies and regulators, pharmaceutical companies, patient representatives, and academics.

Data source identification was restricted to seven European countries where the research team had knowledge of the data landscape and would be able to compare data needs with available data. The selected countries include EU member states from the Northern, Eastern, Southern and Western parts of the continent (Finland, France, Hungary, Italy, Norway, Sweden, United Kingdom).

Methods for case studies

For a set of **selected disease areas (multiple sclerosis, inflammatory bowel disease, lung cancer, and multimorbid patients)**, we assessed the contents of existing data sets in seven European countries for their potential use in answering questions about the efficacy, comparative effectiveness, and safety of new medicines. We used a **case-study approach** to make the exercise feasible. We used a similar approach for all four case study areas, with scope for adjustments in the methods used to allow for disease-specific considerations.

Data source identification

Data sources in the selected countries **(Finland, France, Hungary, Italy, Norway, Sweden, United Kingdom)** were identified for the four disease areas through non-systematic reviews of studies using routine data sources in these countries, online searches and personal knowledge of the research team. The research team was tasked with identifying both disease-specific data sets, such as registries, and generic ones, such as primary care data bases. Rather than compiling a comprehensive list of all available data sources in selected countries, the objective was to identify a range of data sources, representing different types of 'big data' (including routine and administrative data, electronic health records, disease registries) that could be used to support decision-making. High-level information for each data sources was obtained, including whether and which information on demographics, clinical information, outcomes, and resource use was available.

We did not include cohort studies. While these are sometimes labelled 'big data', in particular when including genetic information, they are not typically included in the discourse on the usefulness of existing large data sets for health care system decision-making. While these data sources are 'big', they are typically created for research purposes and therefore leave fewer open questions about their usefulness for establishing efficacy, safety, effectiveness, and other topics regulators and HTA bodies are interested in.

Data needs of health care decision-makers

A case study approach was used in each of the selected disease areas to identify the data needs of regulators and HTA bodies with respect to approval and assessment of new medicines in that therapeutic area. For multiple sclerosis, inflammatory bowel disease and lung cancer, **one drug each** was selected for which assessment reports were available from at least three different HTA agencies (NICE, HAS, TLV). In case several drugs were assessed by different agencies, we selected the one where a bigger role for observational evidence could be expected (e.g. as indicated by a conditional marketing authorisation, or considerable uncertainty regarding the benefit-risk profile at the time of authorisation). For each drug, assessment reports from the European Medicines Agency and HTA bodies in Finland, France, Hungary, Italy, Norway, Sweden and England[1] were searched for and obtained where possible. Researchers then extracted information on the type of evidence that was submitted for regulatory and health technology assessment and recorded uncertainties and open questions mentioned by reviewers.

The approach for the case study on multimorbidities was adjusted due the different nature of the disease. Since multimorbidity is not a condition or therapeutic area in itself, the case study focused on drugs for individual conditions with a high prevalence of co-existing chronic diseases. Three case study drugs, representing three comorbidity disease patterns, were selected and EMA European Public Assessment Reports for recent drugs were screened for indications of taking the effect of the drug on patients with multimorbidities into account.

Comparison of available data with data needs

As a final step, extracted information about regulatory and HTA bodies' data needs was mapped against high-level information on the contents of available data sources.

---

[1] NICE was selected as the only HTA body in the UK. NICE is the national HTA body for England, while separate organisations exist for Wales and Scotland. Data sources were identified UK wide to capture all potentially relevant routine data sources.

2.1.1 Overview of identified data sources

A total of 164 data sources were identified in seven European countries. These comprised of various administrative data bases (such as social insurance data bases, claims data bases, prescription data), death registers (with cause of death), biobanks, surveys, and disease-specific registries. Table 1 gives an overview of the number of data sources identified in each country. Most data sources were generic and could therefore be categorised as 'general' (this category also comprises multimorbidity because multimorbid patients would be included in these data sets, as no specialised multimorbidity registries exist). The numbers in Table 1 relating to multiple sclerosis, inflammatory bowel disease and lung cancer represent dedicated disease registries for these conditions.

Identified data sources were mapped for content to the extent possible with publicly available documentation (see appendix for the full list of data sources and their contents).

TABLE 1: NUMBER OF DATA SOURCES IDENTIFIED BY DISEASE AREA AND COUNTRY

	General/ multimorbidity	Multiple sclerosis	Inflammatory bowel disease	Lung cancer
Finland	10	1	1	1
France	11	1	1	1
Hungary	8	1	1	1
Italy	27	4	2	38
Norway	6	1	-	1
Sweden	6	1	1	1
United Kingdom	23	6	2	7

2.1.2 Overview of case studies

We selected the following drugs as case studies in the four disease areas.

- For **multiple sclerosis**, we focused on alemtuzumab. Alemtuzumab is an interesting case study for assessing the value of routine data, as serious concerns were raised about the safety of the product at the time of approval (with some members of the appraising committee stating their dissent with the approval of the drug). The product is therefore currently placed under additional monitoring, and additional information on safety and effectiveness was requested.
- For **inflammatory bowel disease**, we selected infliximab as case study drug. This was the drug that was appraised by most HTA bodies in this disease area, allowing for more nuanced evaluation of data needs across different institutions.
- For **lung cancer**, we focused on crizotinib. The potential value of non-standard forms of evidence was recognised early by the regulator, as the drug was approved under the conditional marketing authorisation pathway, requiring additional information to be collected in the post-authorisation setting for re-assessment of the product's benefit-risk ratio. The product remains under additional monitoring.
- For **multimorbidities**, case study drugs for three conditions were selected. Selection of conditions was based on clusters of multiple chronic conditions identified in the literature. The body of literature on multimorbidity patterns is growing, with dozens of studies being published over the 20 years. However, some of these only provide associations of chronic conditions, which could randomly occur and therefore not be of relevance to informing decision-making in these diseases. We therefore relied on a systematic review that applied rigorous methodological inclusion and exclusion criteria (Prados-Torres *et al.*, 2014). Each of the three drugs was indicated for a condition that falls into one of three multimorbidity patterns: cardiovascular and metabolic diseases, mental health problems, and musculoskeletal disorders. The case study drug for the first pattern was insulin glargine/lixisenatide combination for treatment of adults with type 2 diabetes mellitus to improve glycaemic control. The case study drug for the second multimorbidity pattern was vortioxetine for treatment of major depressive disorder in adults. Finally, we selected sarilumab for the treatment of moderately to severely active rheumatoid arthritis in adult patients as case study for the third multimorbidity pattern.

2.1.3 Regulatory and HTA data needs summary

Our case studies revealed a variety of different regulatory and HTA information needs. Many of these are typically addressed by randomised controlled trials and might remain unaddressed if these are not conducted – leaving scope for alternative sources of evidence. In our mapping exercise, we aimed to assess whether routine data are available to address these. Specific data needs and how they could be addressed are presented separately for each case study in the appendix. In summary, the following data needs with specific challenges in terms of potential usefulness of routine data were identified.

Safety (adverse events)

Uncertainties in submitted evidence packages often related to safety of new drugs in patients that were excluded from trials in the evidence development programme. This typically included older patients, and special populations, such as pregnant women or patients with renal impairment. To address open questions about the safety of new medicines in terms of adverse events frequency in special populations, hospitalisation data was the most likely valuable administrative data source for serious safety concerns since these events would lead to emergency department visits and/or hospitalisations. However, this information should be linked to data on prescription drug use to identify patients taking the medicine of interest and experiencing an adverse event, which was not always the case in identified data sets. For multimorbid patients, concomitant diagnoses are essential and might not be available from the same data system. Primary care databases appeared more likely to contain comprehensive records on patients' multimorbidities, and sometimes also held additional information such as weight (important for identification of obese patients) and results of laboratory tests (for diagnosis and outcomes).

Efficacy and effectiveness

Similarly to safety, uncertainties about the efficacy of a drug mostly related to patients excluded from trials. An important challenge for assessing efficacy using real-world data is to obtain relevant outcomes data. This is particularly challenging in areas where patient-reported outcomes or other subjective assessments (as opposed to objective measures such as hospital admissions or mortality) are common, as they don't typically feature in routine databases. Such outcomes are common in mental health, and in our case study of vortioxetine, endpoints used in trials were physician-assessed scales and quality of life, for which we could not identify many suitable data sets. Specific scales and patient-reported outcomes were also used as endpoints in trials of other case study drugs. Patient-reported outcomes become increasingly important for establishing efficacy of new medicines, yet we found they are largely missing from routine databases. This represents a gap in big data needs.

Further, some open questions identified in assessments of new medicines required disease-specific outcomes data, such as physician-assessed scales. While these outcomes may not typically feature in general routine data bases, they are more likely to be available from disease-specific registries.

HTA bodies were interested in real-world effectiveness and the comparative effectiveness of new medicines against standard of care, requiring sufficient information about user groups of different medicines to allow comparison of treatment effects. There was also interest in long-term outcomes, requiring tracking of patients over time, with minimal loss to follow-up.

Understanding of drug usage patterns

Both regulators and HTA bodies were interested in better understanding how new medicines are used and by whom. This requires data on treatment adherence. While prescription drug data bases give information on whether prescriptions have been filled, this does not necessarily constitute valid information on treatment adherence. Reasons for discontinuing drugs might only be obtained from more in-depth data bases that can process natural

language input, such as doctors' notes for reasons to switch or discontinue medication.

Safety concerns related to drug-drug interactions that were not tested in the evidence development programme were also highlighted. Real-world data to address this information need includes detailed data on concomitant medication (to identify extent of potential drug-drug interaction) and adverse events.

2.1.4 Limitations

This report and its conclusions should be read in the context of the limitations of the case studies we conducted. We did not exhaustively search and map all potentially relevant data bases in included countries. Rather, we relied on knowledge of researchers in these countries, and on data sources used in previous studies. The identification of potentially relevant data sources was therefore not systematic and the list of data sources should not be regarded as comprehensive.

We assessed the potential usefulness of these data sources based on their content, rather than the quality of the data. Using these data sources for regulatory and HTA decision-making would require an in-depth evaluation of data quality. For example, we did not assess whether validation of the data in these data sources was done, or what the extent of missing data is.

The data sources we identified are at the national or sub-national level. Their potential usefulness for HTA decision-making is therefore likely to be limited to decisions within their respective setting, as HTA bodies might be reluctant to rely on evidence from other countries.

3 Future research priorities

Based on the findings from case studies exploring the potential value of big data for medicines regulatory and HTA decision-making, and a reviews of econometric methods for causal inference in observational big data, a set of recommendations for future research priorities aimed to support the drive towards outcomes-focused, data-based European health care systems has been developed in consultation with the IMI2 DO→IT International Advisory Board.

While some of the findings might be specific to the selected case studies of regulatory and HTA data needs and mapping of available data sources, and the methods review, some overall observations can be made to inform future developments and improve understanding of the future role of big, real-world data in this decision-making setting. In the four case studies, potential contributions of evidence from routine data bases to fill gaps identified in regulatory and HTA reports can be categorised into three main areas: safety (often in relation to populations excluded from pivotal trials), efficacy/effectiveness (again, in relation to real-world populations), and understanding drug use patterns (such as treatment adherence and use of concomitant medication). The case studies also indicate that despite increasing excitement about its potential, observational evidence may not yet have found its place to address regulatory and HTA uncertainties that goes beyond existing uses for these data sources (such as post-authorisation safety monitoring).

3.1.1 Putting big data to use: overcoming barriers and demonstrating value

The future challenge of big data research is not a lack of data sets, but making sure that existing and prospectively collected data is put to use. Greater use requires that data is of high quality, can be easily identified and robustly linked, and that the value of doing so is understood by key stakeholders. There is a need for further research that enables readily identification of existing big data sources and assessment of their quality, sets minimum standards that data collectors can choose to comply with to allow robust linkage at the individual level, and clearly articulates the opportunity costs of not linking data. This should be complemented with clear procedures for obtaining consent for data usage.

Rationale and examples

A first, almost trivial observation is that big, routine data are widely available. A scoping review of data sources in seven European countries from across the continent (Finland, France, Hungary, Italy, Norway, Sweden, UK) revealed a wealth of potentially valuable data sources. Such data sources include primary and secondary care data bases, claims data, surveys, and more specialised data sets such as disease registries.

However, if routine data sources are to be used as a basis for regulatory and HTA decision-making, the information contained in the data sets must be of high quality. This refers to the quality of data collection (use of standard instruments – see ICHOM below), data entry (face validity of records, e.g. absence of ‘upcoding’ or otherwise misleading entries), and completeness of data (lack of systematically missing data). Available data sources should be assessed against quality criteria to ensure the validity and reproducibility of findings. Efforts such as the EMA patient registries initiative can help set standards for data quality in routine data sources that are ‘regulatory-grade’ (EMA, 2017; Miksad and Abernethy, 2018).

The data sets with the greatest potential to inform decision-making in the case studies were those that provided links to other data sets. Linked data (e.g. the English CPRD, or the Italian regional hospitalisation databases) can be used to answer considerably more open questions that regulators or HTA bodies might have than isolated data sets. The notion that data are available but contained in silos has previously been recognised as an important challenge in unlocking the value of big data in health care (Schneeweiss, 2014). Linked data

can support the health care system in a variety of ways, including for improvement of quality of care and health system efficiency (Roski, Bo-Linn and Andrews, 2014; Berger *et al.*, 2015; Salas-Vega, Haimann and Mossialos, 2015; Salcher, 2017). In relation to the latter, the identification of treatments that provide value to patients (and therefore to the health system) can benefit from using large, linked data sets that complement each other. The opportunity costs of not linking data is the health gains foregone by patients not being able to access new treatments that could have been introduced if effectiveness had been evaluated using linked big data.

While primary care data sets appear to be among the richest in terms of baseline information and concomitant diagnoses (therefore a potentially valuable source for covariates necessary to obtain valid treatment effects), these can often lack information on disease-specific outcomes that decision-makers want to see when assessing the clinical value of a medicinal product.

Such disease-specific outcomes can include physician-assessed scales (e.g. tender joint count as important outcome in rheumatoid arthritis) but also patient-reported outcomes (such as pain scales, and instruments that measure overall well-being). Disease-specific outcomes are more likely to be included in specialised disease registries. Creating links between other routine data bases and these registries is therefore important to obtain a complete picture of the patient and could provide decision-makers with the data needed for evidence-based decision-making, including relevant outcomes, and exposure and covariate information, such as medication and concomitant diagnoses.

However, linkage of data sets is far from trivial. Technical and legal challenges exist and overcoming these requires deep knowledge of the legal framework under which two data bases aiming to be linked operate (consideration of data sharing agreements and the scope of informed consent provided by patients included in the data base), and substantial resources to connect the data sets. The harmonisation of data from different sources can pose a significant barrier to increased use of linked data for research. Internationally agreed standards can help make data sets more interoperable in the future, and BD4BO can drive important pioneering work in this area.

3.1.2 Making patient-reported outcomes available in routine data sets

Although a vast number of data sources are available, an important data gap exists for patient-reported outcomes. Patient reported outcomes are becoming increasingly important for the assessment of new treatments but are typically not collected routinely. There is a need to strengthen research into disease-specific minimum sets of patient-reported outcomes that should be collected in routine practice.

Rationale and examples

While a wealth of data is already available in separate routine databases and can used for valuable research when linked together, one notable exception are patient-reported outcomes. Patient-reported outcomes are becoming increasingly important for the evaluation of treatment effects, yet routine databases do not typically collect them. The routine collection of patient-reported outcomes data requires standardisation of the instruments used to measure them. Patient-reported outcome measures (PROMs) exist in many variations, often developed ad hoc and discarding existing instruments measuring similar domains. Efforts to agree on which instruments to use in routine clinical practice are underway (most notably through the International Consortium for Health Outcomes Measurement, ICHOM). The use of PROMs in routine care settings has its own challenges when compared to using PROMs in clinical trials. Feasibility aspects and the burden of data collection on patients and health care professionals need to be taken into account, as routine care

settings will not allow patients and providers to spend considerable time filling out and administering lengthy questionnaires. A potentially valuable source for PROMs in the future could be wearable devices, such as smart phones, watches, etc. Such devices could reduce the administrative burden for collecting patient-reported outcomes. However, it is unclear where the data would be stored, who has access to it, and how it would be linked to other, routine or administrative data sources.

The importance of patient-reported outcomes for regulatory and HTA decision-making is only likely to grow. If big, routine data bases are to become a standard source of evidence for decision-making, ways to incorporate patient-reported outcomes will need to be found. The standardisation of PROMs for clinical practice is therefore an important step in making big data a valuable resource for evaluating efficacy of new medicines.

3.1.3 Increasing acceptability by demonstrating trade-offs

A major necessity for further uptake of big data and non-experimental methods in the regulatory process is cultural change. Research that demonstrates the advantages and limitations of non-experimental methods for causal inference, and identifies situations when non-experimental evidence can complement RCTs, can help build the cultural change necessary to increase the acceptability of non-experimental evidence by regulators and increase the incentives for companies and academics to share their data

Rationale and examples

There is a wealth of existing methods to identify causal effects using observational data. In addition to the well establish methods reviewed in this report, the current development in machine learning also holds great promise for analysis of big data. Still, there are resistance to the use of observational data amongst regulators and HTA bodies. For a cultural change to occur, it is necessary to demonstrate that reliable, robust estimates of treatment effects can be obtained with observational data from routine data sets (such as administrative hospital or primary care data).

The validity of methods to analyse these is essential if they are to be used for regulatory and HTA purposes. Any analysis of observational data faces the problem of dealing with confounders, i.e. factors that affect both whether a patient receives (or chooses) a particular treatment, and the outcome of the treatment. High quality RCTs address this issue by randomly allocating patients to treatment or control group, thereby achieving on average balance in terms of these confounding variables, whether known or unknown, between the two groups.

Novel methods for the analysis of observational data, such as propensity score and other matching methods, and instrumental variable analysis, also aim to achieve balanced treatment and control groups but are limited to controlling for observed confounders (Bosco *et al.*, 2010; Agoritsas *et al.*, 2017). In theory, the availability of more, linked data coming from different sources (and covering not only medical data but also information on other areas, such as lifestyle) to use as covariates could help reduce the risk of bias due to confounding, thus improving internal validity of observational research. Previous efforts to use observational data for regulatory purposes have encountered methodological problems that have somewhat dampened initial excitement about the value of real-world data for post-marketing surveillance (Moore and Furberg, 2015). There remains scope for evaluating whether novel methods for the analysis of observational data can reliably and reproducibly establish causal treatment effects that decision-makers can trust.

The use of big, routine data sources is appealing due its external validity. The issue of trial patients not being representative of real-world populations is of particular relevance for

economic evaluation of new therapies. Modelling of effectiveness of new therapies in older populations with comorbidities is a possible approach to this issue, although they might need to rely on expert opinion to populate models (Guthrie *et al.*, 2012). Large data sets, such as the ones identified in this paper, can provide important insights into the patterns of comorbidities, and how these impact on the effectiveness of medicines, improving the validity of health economic modelling input.

While econometric methods for analysing observational data are potentially increasingly useful as more data can be used to inform models, from a research design perspective, random allocation of patients to treatment and control groups is still widely considered the gold standard for obtaining unbiased treatment effect estimates. The pitfalls of traditional RCTs have been extensively documented, including the lack of representativeness of the actual patient population (Hordijk-Trion *et al.*, 2006). Although randomisation element has been shown to be possible to incorporate into the routine clinical practice settings in so-called pragmatic clinical trials (Ford and Norrie, 2016) as pointed out by e.g. Deaton and Cartwright (2017) no design is stronger than its implementation. Pragmatic trials can be conducted using existing big data infrastructure, such as disease registries and electronic health records (Lund, Oldgren and James, 2017), and can act as a bridge to combine the external validity of valuable real-world data with the internal validity of a methodologically robust study design.

Importantly, there is a need for alignment among data custodians, researchers, policy-makes, and the public about the value of creating large data sets to improve research and inform decision-making for the approval of new medicines. A culture of data sharing and collaboration exists in some countries where linked data appears to be more available (e.g. Swedish registries; UK data linkage initiatives such as CALIBER), but concerted efforts are required to make better use of available data. A recent experience of an initiative to pool available data in Belgium (healthdata.be initiative) shows that active engagement with hundreds of stakeholders is required to create buy-in for the potential of big data to improve health care.

3.1.4 Real-world patients: more to be learned from big data

While there are challenges to including patients reflecting the real-world patient population in traditional clinical trials, pragmatic trials could be used to investigate effectiveness and safety using robust methodological standards in real-world populations as they make use of existing data collection infrastructure while retaining the randomisation process to control for confounding factors.

Rationale and examples

An important finding of the case study in multimorbid patients is that multimorbidity does not feature prominently in the considerations made by the EMA and its review committees. In the three medicines included in the case study, which are indicated for conditions often associated with other chronic conditions (i.e. in multimorbid patients), the EMA reviewers requested post-authorisation studies in two cases to collect information on safety events in patients with characteristics indicating a multimorbid population (e.g. in patients aged 75 years or older). However, lack of evidence on efficacy and safety in multimorbid patients at the time of marketing authorisation did not lead to any restrictions in the approved indication, and open questions regarding the medicines' effect on multimorbid patients identified in this exploratory case study were typically not raised explicitly by the reviewers. Multimorbidity considerations therefore still do not appear to be a priority for the approval of new medicines. Future research on multimorbidities could address some of the reasons for this by improving our understanding of multimorbidity. This could, in turn, also inform regulators and HTA agencies about how to deal with multimorbidities.

First, multimorbidity considerations might not be a primary concern due to lack of information on multimorbidities and patterns. There is still much research to be done to better understand causal pathways between individual diseases of the same multimorbidity pattern. Once such causal pathways are established, regulators will have a strong argument to request evidence on the impact of new medicines on multimorbid patients as part of the evidence development programme. Big data sources are useful for research into multimorbidity patterns and understanding causal relationships. Individual patient IDs allow tracking of patients over time, and integration of data from various source can help construct a complete picture of the patient and his or her interactions with the health care system to better understand how diseases develop. For example, Public Health England are collecting data on all patients with cancer in various formats from GPs, specialists and other health care providers to better understand disease patterns (Rashbass, 2016). More research that uses linked data bases, drawing together complementary information, is needed to understand who the multimorbid patients are that make up the majority of the population treated in health systems, and how their individual conditions relate to each other.

Regulators are tasked with assessing whether new medicines work and are safe. This is most easily done in clinical trials with a homogeneous group of patients. In the evaluation of the clinical evidence submitted for approval of vortioxetine for treatment of major depressive episodes, the reviewers noted that the pivotal trials excluded patients with multimorbidities, but that this was acceptable ‘to reduce confounders and facilitation of evaluation of the pure antidepressant effect’. In addition to methodological considerations, trials including sufficient numbers of multimorbid patients might require a large sample size, potentially significantly increasing the cost of evidence development programmes for new medicines. Nevertheless, decision-makers require evidence on treatment effects in patients reflective of the real world. There is therefore scope for exploring new ways of producing robust evidence, and big, routine data sources have the potential to play an important role in this development by providing the skeleton for pragmatic trials that include real-world patients while achieving high internal validity. While the idea of running pragmatic trials in real-world populations has been around for some time, the increased attention to big data and the potential use of existing data bases for pragmatic trials means that more research into the feasibility of such studies is warranted to identify situations when they are most useful to produce evidence for decision-making.

4 References

Agoritsas, T. *et al.* (2017) 'Adjusted Analyses in Studies Addressing Therapy and Harm: Users' Guides to the Medical Literature', *JAMA*, 317(7), pp. 748–759. doi: 10.1001/jama.2016.20029.

Armstrong, K. (2012) 'Methods in comparative effectiveness research.', *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. American Society of Clinical Oncology, 30(34), pp. 4208–14. doi: 10.1200/JCO.2012.42.2659.

Au Yeung, S. L. *et al.* (2016) 'Genetically predicted 17beta-estradiol, cognitive function and depressive symptoms in women: A Mendelian randomization in the Guangzhou Biobank Cohort Study', *Prev Med*. 2016/04/03. School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, SAR, China. Guangzhou Number 12 Hospital, Guangzhou, China. Department of Public Health and Epidemiology, University of Birmingham, UK. School of Public Heal, 88, pp. 80–85. doi: 10.1016/j.ypmed.2016.03.002.

Augustine, E. F., Adams, H. R. and Mink, J. W. (2013) 'Clinical trials in rare disease: challenges and opportunities', *Journal of child neurology*, 28(9), pp. 1142–1150.

Barnett, K. *et al.* (2012) 'Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study', *The Lancet*, 380(9836), pp. 37–43. doi: 10.1016/S0140-6736(12)60240-2.

Behera, M. *et al.* (2007) 'Evidence-Based Medicine for Rare Diseases: Implications for Data Interpretation and Clinical Trial Design', *Cancer Control*, 14(2), pp. 160–166. doi: 10.1177/107327480701400209.

Berger, M. L. *et al.* (2015) 'Optimizing the leveraging of real-world data to improve the development and use of medicines', *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, 18(1), pp. 127–130. doi: 10.1016/j.jval.2014.10.009.

Bernal, J. L., Cummins, S. and Gasparrini, A. (2017) 'Interrupted time series regression for the evaluation of public health interventions: a tutorial', *International Journal of Epidemiology*, 46(1), pp. 348–355. doi: 10.1093/ije/dyw098.

Bortlik, M. *et al.* (2013) 'Pregnancy and newborn outcome of mothers with inflammatory bowel diseases exposed to anti-TNF-α therapy during pregnancy: three-center study', *Scandinavian Journal of Gastroenterology*, 48(8), pp. 951–958. doi: 10.3109/00365521.2013.812141.

Bosco, J. L. F. *et al.* (2010) 'A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies', *Journal of Clinical Epidemiology*, 63(1), pp. 64–74. doi: 10.1016/j.jclinepi.2009.03.001.

EMA (2017) *Patient Registries Workshop, 28 October 2016: Observations and recommendations arising from the workshop*. London. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2017/02/WC500221618.pdf.

Faria, R. *et al.* (2015) *NICE DSU technical support document 17: The use of observational data to inform estimates of treatment effectiveness in technology appraisal: Methods for comparative individual patient data. Report by the decision support unit*.

Fiorino, G. *et al.* (2017) 'The PROSIT-BIO Cohort', *Inflammatory Bowel Diseases*, 23(2), pp. 233–243. doi: 10.1097/MIB.0000000000000995.

Ford, I. and Norrie, J. (2016) 'Pragmatic Trials', *New England Journal of Medicine*. Edited by J. M. Drazen et al. Massachusetts Medical Society, 375(5), pp. 454–463. doi: 10.1056/NEJMra1510059.

Franklin, J. M. and Schneeweiss, S. (2017) 'When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials?', *Clinical Pharmacology & Therapeutics*, 102(6), pp. 924–933. doi: 10.1002/cpt.857.

Fretheim, A. *et al.* (2013) 'Interrupted time-series analysis yielded an effect estimate concordant with the cluster-randomized controlled trial result', *J Clin Epidemiol*. 2013/07/03. Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health

Care Institute, Boston, MA 02215, USA. atle.fretheim@nokc.no, 66(8), pp. 883–887. doi: 10.1016/j.jclinepi.2013.03.016.
Fretheim, A. *et al.* (2015) 'A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation', *J Clin Epidemiol*. 2014/12/17. Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 133 Brookline Avenue, Boston, MA 02215, USA; Global Health Unit, Norwegian Knowledge Centre for the Health Services, PO Box 7004, St. Olavs pl, 0130 Oslo, 68(3), pp. 324–333. doi: 10.1016/j.jclinepi.2014.10.003.
Gallin, J. I. *et al.* (2003) 'Itraconazole to prevent fungal infections in chronic granulomatous disease', *New England Journal of Medicine*, 348(24), pp. 2416–2422.
Groves, P. *et al.* (2013) *The 'big data' revolution in healthcare: Accelerating value and innovation*. Available at: http://repositorio.colciencias.gov.co:8081/jspui/bitstream/11146/465/1/1661-The_big_data_revolution_in_healthcare.pdf.
Gustavsson, A. *et al.* (2007) 'A 2-year follow-up of the Swedish-Danish infliximab/placebo trial in steroid resistant acute ulcerative colitis', *Gastroenterology*, 132(4 Suppl 2), pp. A146-7.
Guthrie, B. *et al.* (2012) 'Adapting clinical guidelines to take account of multimorbidity.', *BMJ (Clinical research ed.)*. British Medical Journal Publishing Group, 345, p. e6341. doi: 10.1136/BMJ.E6341.
Habl, C. *et al.* (2016) *Study on Big Data in Public Health, Telemedicine and Healthcare*. Brussels, Belgium. Available at: https://ec.europa.eu/health/sites/health/files/ehealth/docs/bigdata_report_en.pdf (Accessed: 24 January 2017).
Halpin, S. J. *et al.* (2013) 'Efficacy of infliximab in acute severe ulcerative colitis: a single-centre experience.', *World journal of gastroenterology*. Baishideng Publishing Group Inc, 19(7), pp. 1091–7. doi: 10.3748/wjg.v19.i7.1091.
Hanauer, S. B. *et al.* (2004) 'Incidence and importance of antibody responses to infliximab after maintenance or episodic treatment in Crohn's disease.', *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*, 2(7), pp. 542–53. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15224278 (Accessed: 5 March 2018).
Harris, A. D. *et al.* (2006) 'The use and interpretation of quasi-experimental studies in medical informatics', *Journal of the American Medical Informatics Association*, 13(1), pp. 16–23.
Hordijk-Trion, M. *et al.* (2006) 'Patients enrolled in coronary intervention trials are not representative of patients in clinical practice: results from the Euro Heart Survey on Coronary Revascularization', *European Heart Journal*, 27(6), pp. 671–678. doi: 10.1093/eurheartj/ehi731.
Jadad, A. R. and Enkin, M. W. (2007) *Randomized Controlled Trials: Questions, Answers and Musings*. 2nd edn. London: Wiley & Sons. Available at: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1405132663.html.
Järnerot, G. *et al.* (2005) 'Infliximab as rescue therapy in severe to moderately severe ulcerative colitis: a randomized, placebo-controlled study.', *Gastroenterology*, 128(7), pp. 1805–11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15940615 (Accessed: 5 March 2018).
Jonker, C. J. *et al.* (2017) 'Registries supporting new drug applications', *Pharmacoepidemiology and Drug Safety*, p. n/a-n/a. doi: 10.1002/pds.4332.
Katz, J. A. *et al.* (2004) 'Outcome of pregnancy in women receiving infliximab for the treatment of Crohn's disease and rheumatoid arthritis.', *The American journal of gastroenterology*, 99(12), pp. 2385–92. doi: 10.1111/j.1572-0241.2004.30186.x.
Khozin, S., Blumenthal, G. M. and Pazdur, R. (2017) 'Real-world Data for Clinical Evidence Generation in Oncology', *JNCI: Journal of the National Cancer Institute*, 109(11), p. djx187.
van Leeuwen, N. *et al.* (2016) 'Regression Discontinuity Design: Simulation and Application in Two Cardiovascular Trials with Continuous Outcomes', *Epidemiology*. 2016/04/01. From

the aCentre for Medical Decision Making, Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands; bDepartment of Gerontology and Geriatrics, Leiden University Medical Centre, Leiden, The Netherlands; cInstitute for Evidence-Ba, 27(4), pp. 503–511. doi: 10.1097/ede.0000000000000486.
Lund, L. H., Oldgren, J. and James, S. (2017) 'Registry-Based Pragmatic Trials in Heart Failure: Current Experience and Future Directions.', *Current heart failure reports*. Springer, 14(2), pp. 59–70. doi: 10.1007/s11897-017-0325-0.
Lutz, W. *et al.* (2016) 'Clinical effectiveness of cognitive behavioral therapy for depression in routine care: A propensity score based comparison between randomized controlled trials and clinical practice', *J Affect Disord*. 2015/10/05. Clinical Psychology and Psychotherapy, Department of Psychology, University of Trier, D-54296 Trier, Germany. Electronic address: wolfgang.lutz@uni-trier.de. Clinical Psychology and Psychotherapy, Department of Psychology, University of Trier, D-54296 Tri, 189, pp. 150–158. doi: 10.1016/j.jad.2015.08.072.
Maas, I. L. *et al.* (2017) 'The regression discontinuity design showed to be a valid alternative to a randomized controlled trial for estimating treatment effects', *J Clin Epidemiol*. 2016/11/21. Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charite - Universitatsmedizin Berlin, Chariteplatz 1, 10117 Berlin, Germany. Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charite - , 82, pp. 94–102. doi: 10.1016/j.jclinepi.2016.11.008.
Mayne, S. L., Lee, B. K. and Auchincloss, A. H. (2015) 'Evaluating propensity score methods in a quasi-experimental study of the impact of menu-labeling', *PloS one*, 10(12), p. e0144962.
Miksad, R. A. and Abernethy, A. P. (2018) 'Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality', *Clinical Pharmacology & Therapeutics*, 103(2), pp. 202–205. doi: 10.1002/cpt.946.
Moore, T. J. and Furberg, C. D. (2015) 'Electronic Health Data for Postmarket Surveillance: A Vision Not Realized', *Drug Safety*, 38(7), pp. 601–610. doi: 10.1007/s40264-015-0305-9.
Moscoe, E., Bor, J. and Bärnighausen, T. (2015) 'Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice', *Journal of Clinical Epidemiology*, 68(2), pp. 132–143. doi: https://doi.org/10.1016/j.jclinepi.2014.06.021.
Murdoch, T. B. and Detsky, A. S. (2013) 'The Inevitable Application of Big Data to Health Care', *JAMA*, 309(13), pp. 1351–1352. doi: 10.1001/jama.2013.393.
Nichol, A. D. *et al.* (2010) 'Challenging issues in randomised controlled trials', *Injury*, 41, pp. S20–S23. doi: 10.1016/j.injury.2010.03.033.
Park, W. *et al.* (2017) 'Efficacy and safety of switching from reference infliximab to CT-P13 compared with maintenance of CT-P13 in ankylosing spondylitis: 102-week data from the PLANETAS extension study', *Annals of the Rheumatic Diseases*, 76(2), pp. 346–354. doi: 10.1136/annrheumdis-2015-208783.
Prados-Torres, A. *et al.* (2014) 'Multimorbidity patterns: a systematic review', *Journal of Clinical Epidemiology*. Pergamon, 67(3), pp. 254–266. doi: 10.1016/J.JCLINEPI.2013.09.021.
Rashbass, J. (2016) *New cancer dashboards - So much data; enough information?*, *Public health matters blog (Public Health England)*. Available at: https://publichealthmatters.blog.gov.uk/2016/05/13/new-cancer-dashboards-so-much-data-enough-information/ (Accessed: 5 February 2018).
Relton, C. L. and Davey Smith, G. (2012) 'Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease', *International journal of epidemiology*, 41(1), pp. 161–176.
Roski, J., Bo-Linn, G. W. and Andrews, T. A. (2014) 'Creating value in health care through big data: opportunities and policy implications', *Health Affairs (Project Hope)*, 33(7), pp. 1115–1122. doi: 10.1377/hlthaff.2014.0147.
Rutgeerts, P. *et al.* (2004) 'Comparison of scheduled and episodic treatment strategies of infliximab in Crohn's disease.', *Gastroenterology*, 126(2), pp. 402–13. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/14762776 (Accessed: 5 March 2018).
Salas-Vega, S., Haimann, A. and Mossialos, E. (2015) 'Big Data and Health Care: Challenges and Opportunities for Coordinated Policy Development in the EU', *Health Systems & Reform*, 1(4), pp. 285–300. doi: 10.1080/23288604.2015.1091538.
Salcher, M. (2017) 'Connecting the dots: putting big data to work for health systems', *Eurohealth*, 23(1). Available at: http://www.lse.ac.uk/LSEHealthAndSocialCare/Outputs/eurohealth/eurohealth.aspx (Accessed: 25 May 2017).
Sandborn, W. *et al.* (2008) 'SONIC: A randomized, double-blind, controlled trial comparing Infliximab and Infliximab plus Azathioprine to Azathioprine in patients with Crohn's Disease naive to', *American Journal of Gastroenterology*, 103, p. 1117. Available at: https://journals.lww.com/ibdjournal/Citation/2008/12003/SONIC__A_randomized,_double_blind,_controlled.1.aspx (Accessed: 5 March 2018).
Sandborn, W. J. *et al.* (2009) 'Colectomy Rate Comparison After Treatment of Ulcerative Colitis With Placebo or Infliximab', *Gastroenterology*, 137(4), pp. 1250–1260. doi: 10.1053/j.gastro.2009.06.061.
Sands, B. E. *et al.* (2004) 'Infliximab maintenance therapy for fistulizing Crohn's disease.', *The New England journal of medicine*, 350(9), pp. 876–85. doi: 10.1056/NEJMoa030815.
Schmidt, A. F. *et al.* (2016) 'Adjusting for Confounding in Early Postlaunch Settings: Going Beyond Logistic Regression Models', *Epidemiology (Cambridge, Mass.)*, 27(1), pp. 133–142. doi: 10.1097/EDE.0000000000000388.
Schneeweiss, S. (2006) 'Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics', *Pharmacoepidemiology and Drug Safety*, 15(5), pp. 291–303. doi: 10.1002/pds.1200.
Schneeweiss, S. (2014) 'Learning from Big Health Care Data', *New England Journal of Medicine*, 370(23), pp. 2161–2163. doi: 10.1056/NEJMp1401111.
Schnitzler, F. *et al.* (2011) 'Outcome of pregnancy in women with inflammatory bowel disease treated with antitumor necrosis factor therapy.', *Inflammatory bowel diseases*, 17(9), pp. 1846–54. doi: 10.1002/ibd.21583.
Schweizer, M. L., Braun, B. I. and Milstone, A. M. (2016) 'Research methods in healthcare epidemiology and antimicrobial stewardship—quasi-experimental designs', *infection control & hospital epidemiology*, 37(10), pp. 1135–1140.
Sedgwick, P. (2012) 'External and internal validity in clinical trials', *BMJ*. British Medical Journal Publishing Group, 344(feb16 1), pp. e1004–e1004. doi: 10.1136/bmj.e1004.
Soumerai, S. B. and Koppel, R. (2017) 'The reliability of instrumental variables in health care effectiveness research: Less is more', *Health services research*, 52(1), pp. 9–15.
Streeter, A. J. *et al.* (2017) 'Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review', *Journal of clinical epidemiology*, 87, pp. 23–34.
Weber, G. M., Mandl, K. D. and Kohane, I. S. (2014) 'Finding the Missing Link for Big Biomedical Data', *JAMA*, 311(24), pp. 2479–2480. doi: 10.1001/jama.2014.4228.
West, S. G. *et al.* (2008) 'Alternatives to the randomized controlled trial', *American journal of public health*, 98(8), pp. 1359–1366.
Wilhelm, S. M. *et al.* (2008) 'A review of infliximab use in ulcerative colitis', *Clinical Therapeutics*, 30(2), pp. 223–230. doi: 10.1016/j.clinthera.2008.02.014.
Yoo, D. H. *et al.* (2016) 'A phase III randomized study to evaluate the efficacy and safety of CT-P13 compared with reference infliximab in patients with active rheumatoid arthritis: 54-week results from the PLANETRA study.', *Arthritis research & therapy*, 18(1), p. 82. doi: 10.1186/s13075-016-0981-6.
Zhou, H. *et al.* (2016) 'Difference-in-Differences method in Comparative effectiveness research: Utility with Unbalanced groups', *Applied health economics and health policy*, 14(4), pp. 419–429.

Appendix

Contents:

1. Case study I: Multiple sclerosis - Potential value of big data for the approval of alemtuzumab
2. Case study II: Inflammatory bowel disease - Value of big data for the approval of infliximab
3. Case study III: Lung cancer - Potential value of big data for the approval of crizotinib
4. Case study IV: Multimorbid patients - Potential value of big data for the approval of medicines in chronic conditions
5. Full report: review of methods to analyse big data
6. Big data sources in 7 European countries

Case study I

Multiple sclerosis

Potential value of big data for the approval of alemtuzumab

Case study I
Multiple sclerosis
Potential value of big data for the approval of alemtuzumab

Multiple sclerosis (MS) is a chronic autoimmune disease that affects the central nervous system. The disease results in damaged myelin, a protective layer of protein that covers nerve fibres in the brain and spinal cord, and axons, the underlying nerve fibre. This damage can in turn result in significant disabilities.
As a progressive disease, treatments for MS are used to limit progression of the disease or to treat symptoms. One such treatment is alemtuzumab (trade name Lemtrada) which works by regulating the immune system.

Data sources

Routine data sources for MS are available in each of the seven countries covered in the scoping review. Nearly all of these are generic data sets, covering many more diseases than MS alone. A smaller number of data sources are disease-specific, focusing only on MS. Nearly all identified MS-specific data sources are registries at either a national or local level. In the United Kingdom, in addition to registries, a survey and research study were also identified. The following table summarises the data sources that were identified. Further details on all identified data sources are available in appendix 1.

TABLE 2: IDENTIFIED DATA SOURCES RELEVANT TO MULTIPLE SCLEROSIS RESEARCH IN SEVEN EUROPEAN COUNTRIES

	General (thought to include MS among other diseases)	MS-specific data source	Name of MS-specific data source
Finland	10	1	MS disease registry
France	11	1	French Multiple Sclerosis Registry (OFSEP)
Hungary	8	1	Csongràd County MS registry; multiple sclerosis centres with a disease registry (19)
Italy	27	4	National registry; Regional registry (Tuscany); Regional registry (Liguria); Regional registry (Sicily)
Norway	6	1	National quality register for Multipel Skleros
Sweden	6	1	National quality register for Multiple Sclerosis (SMSreg)
United Kingdom	23	6	Registry (4), survey (1), research study (1)

Regulatory and HTA data needs

Pharmaceutical assessment reports were analysed to determine:

- how the reports use data sources, and
- which data is needed by the reports.

Both European and national-level assessment reports were examined. This research focused on one drug, Lemtrada, which was chosen based on the availability of assessment reports from the majority of countries in this study. When assessment reports could not be accessed for a country, they were excluded.
The EMA's European Public Assessment Report (EPAR) was published in June 2013 and was followed by national assessment reports published between February 2014 and January 2016. Ranging in length from two pages to 116 pages, the reports examined

Lemtrada for adult patients with relapsing-remitting multiple sclerosis. The following table introduces the assessment reports that were examined.

TABLE 3: INCLUDED ASSESSMENT REPORTS ON LEMTRADA (ALEMTUZUMAB)

	Assessment report
Europe	EUROPEAN PUBLIC ASSESSMENT REPORT Lemtrada International non-proprietary name: ALEMTUZUMAB Procedure No. EMEA/H/C/003718/0000
Finland	Excluded
France	HAS BRIEF SUMMARY OF THE TRANSPARENCY COMMITTEE OPINION - Lemtrada
Hungary	Excluded
Italy	ANNEX I - SUMMARY OF PRODUCT CHARACTERISTICS
Norway	Excluded
Sweden	TLV Lemtrada (alemtuzumab) Health economics knowledge base Evaluated indication: Treatment of adult patients with forest-induced multiple sclerosis (RRMS) with active disease defined by clinical findings or image findings.
England	NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE Final appraisal determination Alemtuzumab for treating relapsing–remitting multiple sclerosis

How have the reports used data?

As expected, clinical trials data is the most prevalent form of data used in the assessment reports. All five reports mentioned clinical trials data. Two out of five reports used different types of data in addition to that obtained through clinical trials. For example, the English body, NICE, used a range of data sources to inform their economic model, and expanded on the trial data by submitting a meta-analysis of trials and a comparison of alemtuzumab with other treatments for active relapsing-remitting multiple sclerosis. Sweden used a Markov model and network analysis.

How have the reports used real-world data?

Real-world data was rarely used in the assessment reports, and when present was used in a limited way. For example, the NICE report described a matrix that was used by the manufacturer to represent the natural history transition and disability progression in people without therapies. This matrix was based in part on the London Ontario dataset, which is a longitudinal observational study. Health state utility values were also used in the NICE report, collected from a UK survey of health-related quality of life (EQ-5D) in people with multiple sclerosis (Orme et al, 2007).

What data needs have been identified by the reports?

Most assessment reports noted data that was missing or that they would like to have been included. This tended to be data related to the safety or effectiveness of the drug, as well as information about the long-term effects of the drug, data that can support clinical trials, comparative data, and additional requirements such as periodic safety update reports.
For example, the European report cited the following information as missing or limited, all related to safety:

- Impact on fertility
- Use during pregnancy
- Use during lactation
- Paediatric use
- Use in patients aged >55 years (including use in elderly patients aged ≥65 years)
- Impact on response to vaccination and value of pre-treatment vaccination
- Use in patients with renal impairment

- Use in patients with hepatic impairment
- Use in patients with human immunodeficiency virus(HIV)
- Use in patients with Hepatitis B virus (HBV)
- Use in patients with Hepatitis C virus (HCV)

Similar missing information was echoed in a number of other assessment reports as well. For example, the Italian report notes the lack of data on the association of the drug with the reactivation of the hepatitis B or C virus, noting that this is because patients with evidence of such infections are excluded from clinical trials. The report also notes that:

- The safety of vaccinations is unknown after treatment Lemtrada, laboratory analysis is recommended to monitor patients for signs of autoimmune diseases.
- It is difficult to estimate the frequency of a causal link with some adverse reactions with alemtuzumab.
- There was no formal drug interaction for Lemtrada at the dose recommended for patients with MS.
- There is limited data on Lemtrada for pregnant women.
- It is unknown whether alemtuzumab is excreted in breast milk so the benefits versus risk of breast feeding is not known.
- There is no data on the safety on Lemtrada in relation to fertility.
- There is no data on how Lemtrada could affect a person's ability to drive or to use machines.
- There is no data on whether alemtuzumab is carcinogenic or matagenic.
- It is not known whether or not senior citizens respond differently to the drug than younger patients.

Others, including Sweden and France, were concerned with a lack of comparative data. Problems with RCTs were raised. For example, the NICE report mentioned that there were inconsistencies in the definitions and populations of subgroups in clinical trials and that the mixed treatment comparison depended heavily on indirect evidence.

The Swedish report writes favourably about the ability of models to predict cost and effects beyond the end of clinical studies, suggesting that there may be some openness to other techniques – perhaps using real-world data – that could achieve similar aims. Similarly, the NICE report mentioned that due to an absence of long-term data the long-term benefit of alemtuzumab is unknown, suggesting that there could be some appetite for this. The NICE report also notes that there was no data to support the assumption of "constant treatment effect throughout the course of a person's multiple sclerosis up to EDSS state 7 or secondary progressive multiple sclerosis".

At least two reports outlined additional requirements they wanted. In Italy's case, periodic safety update reports were requested, as was a risk management plan, and an education program for healthcare professionals and patients containing information on the risks associated with Lemtrada and how to reduce those risks. The European report called for:

- Ongoing and planned studies in the PhV development plan
- Characterising the long-term safety profile of alemtuzumab in patients with RRMS in a real-world setting
- Assessing the effectiveness of risk minimisation measures
- Assessing adverse pregnancy outcomes in women exposed to alemtuzumab
- Evaluate the efficacy, safety and tolerability of alemtuzumab versus appropriate comparator in paediatric subjects with relapsing forms of MS who have disease activity on prior therapy
- Assessing use in patients with human immunodeficiency virus (HIV)

Data gaps: mapping of available data sources to regulators' and HTA bodies' data needs

Once data needs had been identified based on assessment reports, these were matched with data sources that could produce added value and inform decision-making. The following table summarises the data needs and potential uses for the data sources. It indicates how we might begin to think about using real-world data to support decision-making among HTA bodies and regulators.

TABLE 4: OPEN QUESTIONS ON ALEMTUZUMAB FOR TREATMENT OF MULTIPLE SCLEROSIS AND AVAILABLE DATA SOURCES IN THE UK

Data need	How could routine data be used	Potential data sources and methods
Safety and effectiveness of the drug	As clinical trials exclude certain populations and operate in an idealised environment, there is often limited information about certain groups of people. The MS reports highlight paediatrics, seniors, pregnant women, and people with Hepatitis B or C among others. There is also interest in understanding the effect of the drug in combination with other treatments, for example vaccines.	CPRD, SAIL, HES, ONS mortality data, UK biobank (link to HES), MSbase, EUreMS, Scottish MS Register, EQ-5D and other surveys. The data source used will depend on the outcome of interest. For example, if regulators/HTA require QoL information then a registry would be most suitable.
Long-term effects of the drug	There is some appetite to understand the effects of a drug beyond the scope of a clinical trial. This can be achieved through predictive modelling as well as real-world data. For example, RWD can contribute to understanding quality of life, the history of MS, and adherence to treatment.	CPRD, SAIL, HES, ONS mortality data, UK biobank (link to HES), MSbase, EUreMS, Scottish MS Register, EQ-5D and other surveys, Labour Force Survey, Millennium Cohort Study
Data to support modelling	Economic modelling relies on accurate datasets and good assumptions. RWD can support these models and inform the model's parameters.	CPRD, SAIL, HES, ONS mortality data, UK biobank (link to HES), MSbase, EUreMS, Scottish MS Register, EQ-5D and other surveys
Comparative data	There is a desire to compare Lemtrada with comparators, sometimes among certain population sub-groups. RWD can provide information about the drug and a comparator and assist in choosing appropriate comparators.	CPRD, SAIL, HES, ONS mortality data, UK biobank (link to HES), MSbase, EUreMS, Scottish MS Register, EQ-5D and other surveys
Data to support additional requirements	There is interest in obtaining additional information after the assessment report including periodic safety update reports, risk management plans, and education programs for patients and practitioners. RWD can provide some of this information that may be beyond the scope of RCTs.	CPRD, SAIL, HES, ONS mortality data, UK biobank (link to HES), MSbase, EUreMS, Scottish MS Register, EQ-5D and other surveys, Labour Force Survey, Millennium Cohort Study

Case study II

Inflammatory bowel disease

Potential value of big data for the approval of infliximab

Case study II
Inflammatory bowel disease
Potential value of big data for the approval of infliximab

Infliximab is a chimeric monoclonal antibody that is used to treat several chronic inflammatory diseases, including Inflammatory Bowel Diseases, IBD, (Crohn's disease and ulcerative colitis), rheumatoid arthritis, psoriasis, psoriatic arthritis and ankylosing spondylitis. Infliximab is a necrosis factor alpha (TNFα) inhibitor and works by binding (with high affinity) to both soluble and transmembrane forms of TNFα, a protein produced by the body which has an important role in promoting inflammation. Infliximab is administrated by intravenous (IV) infusion (Reference: EMA reports).
Brand names for infliximab include Remicade, Inflectra, Flixabi, and Remsima. Remicade was first approved by the European Medicines Agency (EMA) in August, 1999. The other brand names listed are biosimilars for Remicade and, in general, have been approved over the last 8-9 years by EMA and then approved in varying degrees in single-member nations. Indications for Infliximab related to IBD include adult Crohn's disease (CD), paediatric Crohn's disease and ulcerative colitis (UC). Other indications for infliximab include rheumatoid arthritis (RA), ankylosing spondylitis (AP), psoriasis, and psoriatic arthritis (PA). Details for the indications for IBD are provided below.

BOX 1: OVERVIEW OF INDICATIONS FOR INFLIXIMAB

Adult Crohn's disease
• treatment of moderately to severely active Crohn's disease, generally in adult patients who have not responded despite a full and adequate course of therapy with a corticosteroid and/or an immunosuppressant; or who are intolerant to or have medical contraindications for such therapies.
• treatment of fistulising, active Crohn's disease, in adult patients who have not responded despite a full and adequate course of therapy with conventional treatment (including antibiotics, drainage and immunosuppressive therapy).
Paediatric Crohn's disease
• indicated for treatment of severe, active Crohn's disease, in children and adolescents aged 6 to 17 years, who have not responded to conventional therapy including a corticosteroid, an immunomodulator and primary nutrition therapy; or who are intolerant to or have contraindications for such therapies.
Ulcerative colitis
• indicated for treatment of moderately to severely active ulcerative colitis in adult patients who have had an inadequate response to conventional therapy including corticosteroids and 6-mercaptopurine (6-MP) or azathioprine (AZA), or who are intolerant to or have medical contraindications for such therapies. Infliximab has been studied only in combination with conventional immunosuppressive therapy.

Data sources

Data regarding the evidence used in the approval process of Infliximab in Europe has been limited to reviews of the scientific literature from randomised, controlled clinical trials (RCTs). No evidence of alternative data use in the approval process has emerged for this drug, which is centrally approved by the EMA for all countries surveyed. There is some evidence of approval processes and HTA reports in single countries, but again, they appear to have only collected evidence of clinical efficacy from RCTs, with some use of observational studies for safety surveillance to update recommendations.

An overview of the evidence consulted during the approval process is provided below, with details regarding specific clinical trials and subsequent studies provided below.
From the EMA reports,
http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/000240/human_med_001023.jsp&mid=WC0b01ac058001d124):

- For Crohn's disease, Remicade has been compared with placebo (a dummy treatment) in 1,090 adults in four studies. The main measures of effectiveness were the improvement in the severity of symptoms or the healing of fistulae. The effects of adding Remicade to existing treatment have also been studied in 103 children and adolescents with Crohn's disease who were aged between six and 17 years. A sixth study in 508 adult patients looked at the number of patients whose symptoms improved and who did not need additional treatment with corticosteroids (other medicines used in Crohn's disease). The patients were treated for six months with Remicade, another medicine azathioprine, or the combination of Remicade and azathioprine.
- For ulcerative colitis (728 adults), ankylosing spondylitis (70 adults), psoriatic arthritis (104 adults) and psoriasis (627 adults), Remicade has been compared with placebo. In all of these studies, the main measure of effectiveness was the change in symptoms after up to 16 weeks.
- For ulcerative colitis, Remicade has also been studied in 60 children aged between six and 17 years. The main measure of effectiveness was the number of patients who responded to treatment at week eight, after having received three infusions with Remicade.

Randomised clinical trial data has been cited in all documentation from the EMA (and for the FDA in the United States) to initial approval of Infliximab in the treatment of IBD (see Box 2). Two important clinical trials for Crohn's Disease (CD) were ACCENT 1, a large, multicentre trial, which showed efficacy in maintaining remission in inflammatory CD and ACCENT 2, another large Phase III clinical trial, which showed infliximab to be beneficial in maintain closure of fistulae (Hanauer *et al.*, 2004; Rutgeerts *et al.*, 2004; Sands *et al.*, 2004). Evidence supporting the treatment of ulcerative colitis (UC) with infliximab for improving health-related quality of life (HRQOL) date from 2005 in the Acute ulcerative Colitis Treatment trials (ACT1 and ACT2). (Wilhelm *et al.*, 2008)

Various RCT studies followed to further test Infliximab in IBD (Järnerot *et al.*, 2005; Gustavsson *et al.*, 2007), including the RESULTS UC study (Sandborn *et al.*, 2009) and the SONIC study (Sandborn *et al.*, 2008). There have been observational studies, registry studies and systematic reviews to determine safety for use in pregnancy (Katz *et al.*, 2004; Schnitzler *et al.*, 2011; Bortlik *et al.*, 2013). Studies using real-life clinical practice to test effectiveness have also been observed (Halpin *et al.*, 2013).

BOX 2: CLINICAL EVIDENCE FOR THE EFFICACY AND SAFETY OF INFLIXIMAB IN IBD (SOURCE: EMA ASSESSMENT HISTORY DOCUMENTS FOR REMICADE).

Table 1 Overview of studies in support of efficacy and safety

Clinical Trial	Phase	Patient Population	Number of patients
0168T08	I	Active Crohn's disease	10
0168T11	II	Moderate to severe Crohn's disease	21
0168T16	II/III	Moderate to severe Crohn's disease	108
0168T20	II/III	Fistulising Crohn's disease	94
0168T24[1]	III	Moderate to severe Crohn's disease	40
0168T07	I	Active RA	20
0168T09	II	Active RA	73
0168T14	II	Active RA	101
0168T15/T17[2]	II	Active RA	28
0168T18	I	Active RA	16
0168T22	III	Active RA	428
0168T03	I/II	Healthy volunteers	39
0168T00	I	Compassionate use	9
0168T12	II	Ulcerative colitis	11

1: Open label study in patients earlier included in studies T 08, T11, T16 and T20.
2: These two trials had separate protocols, but included the same patients and were reported in a single study report

Further data with regard to long-term treatment of active CD (trial ACCENT I; C0168T21) and fistulising CD (ACCENT II; C0168T26) were submitted later on. ACCENT I was a multicentre, randomised, double-blind, clinical trial of maintenance infliximab treatment compared with a single dose of infliximab in 580 patients with moderately to severely active Crohn´s disease (CD). ACCENT II was a multicentre, randomised, double-blind, clinical trial of maintenance infliximab treatment compared with a 3-dose induction regimen of infliximab only in 306 patients with fistulising Crohn´s disease.

- Study C0168T37 (ACT 1): 54-week study of 364 subjects, designed as a randomised, placebo-controlled, double blind trial to evaluate the safety and efficacy of infliximab in patients with active UC.

- Study C0168T46 (ACT 2): 30-week study of 364 subjects designed as a randomised, placebo-controlled, double blind trial to evaluate the safety and efficacy of infliximab in patients with active UC.

Source: EMA assessment history documents for Remicade (http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/000240/human_med_001023.jsp&mid=WC0b01ac058001d124, accessed January 2018).

Biosimilars:

- The EMA approval of the biosimilars for Remicade followed the standard requirements for such products, that is,
    - "The Agency's Committee for Medicinal Products for Human Use (CHMP) decided that, in accordance with EU requirements, Inflectra (Flixabi, Remsima) has been shown to have a comparable quality, safety and efficacy profile to Remicade. Therefore, the CHMP's view was that, as for Remicade, the benefit outweighs the identified risks. The Committee recommended that Inflectra (Flixabi, Remsima) be approved for use in the EU." (source: EMA documentation for each biosimilar for Remicade).

Real-world data Sources

IBD has been studied using routinely collected, general data sources in all the countries surveyed in this report, however, few disease-specific data sources could be identified (see table below). Only Sweden and the UK have national disease registries for IBD patients, while in Hungary, a paediatric IBD registry is listed, though it appears to be quite limited (only 44 gastroenterologists are involved). With difficulty, two locally-managed IBD registries were identified in Italy.

HTA assessments and data needs

The following HTA and guidance reports were consulted to identify data needs and data sources. As outlined above, the HTA reports were based on clinical efficacy data, the same used during the EMA approval process and documented in the assessment history cited on their website.

HTA reports and documents

- HEALTH TECHNOLOGY ASSESSMENT, VOLUME 20 ISSUE 39 MAY 2016, ISSN 1366-5278, Infliximab, adalimumab and golimumab for treating moderately to severely active ulcerative colitis after the failure of conventional therapy (including a review of TA140 and TA262): clinical effectiveness systematic review and economic model. Authors: Rachel Archer, Paul Tappenden, Shijie Ren, Marrissa Martyn-St James, Rebecca Harvey, Hasan Basarir, John Stevens, Christopher Carroll, Anna Cantrell, Alan Lobo and Sami Hoque.
- NICE Guidance on Infliximab, adalimumab and golimumab for treating moderately to severely active ulcerative colitis after the failure of conventional therapy. 2015.
- HAS Haute Autorité de Santé – Inflectra, Remsima, Flixabi report finding that the biosimilars were approved for clinical use in the same manner and for the same indications as Remicade.
- Canadian Agency for Drugs and Technologies in Health. HTA Issue 120, July 2009. Anti_TNF-α Drugs for Refactory Inflammatory Bowel Disease: Clinical and Cost-Effectiveness Analysis.
- Approval documents for the Italian regions of Emilia-Romagna and Veneto and a Swedish agency, Tandvards-OCH Lakemedelsformansverket (TLV).

TABLE 5: IBD DATA SOURCES

	General	IBD-specific data source	Name of IBD-specific data source
Finland	10	NA	No IBD disease registry could be identified, but the Finnish Special Reimbursements for Drug Costs Registry has been used to identify IBD patients for observational studies
France	11	1	There is no specific IBD registry, but the national, public French cohort of Inflammatory Bowel Disease (IBD) is used for research purposes.
Hungary	8	1	HUPIR (Hungarian Pediatric IBD Registry) is a national registry for paediatric patients managed by 44 paediatric gastroenterologists in Hungary
Italy	27	2	2 IBD disease registries (one regional – Liguria – and one provincial - Forlì)
Norway	6	NA	No IBD disease registry could be identified, but the Norwegian Patient Registry for hospital care has been used to identify IBD patients for observational studies
Sweden	6	1	National quality register for Inflammatory Bowel Disease
United Kingdom	23	1	National Register of IBD patients of the British Society of Gastroenterology

What data needs have been identified by the reports?

EMA report (Summary of product characteristics for Remicade) – Current version

When Remicade was originally approved as the first commercially-distributed form of Infliximab, a number of gaps in information were identified. However, the drug has been on the market since 1999, and data has been collected for many of the safety issues associated with short- and long-term use. The issues which still require additional information in the report include:

- Concurrent administration with other biological therapeutics
- The incidence of delayed hypersensitivity reactions after Remicade-free intervals of more than 1 year
- The effects on fertility and general reproductive function. Though the report lists evidence from observational studies of the use of infliximab during pregnancy, it still states that the "available clinical experience is limited" and that it should be used in pregnancy only if clearly needed.
- Breastfeeding: it is unknown whether infliximab is excreted in human milk or absorbed systemically after ingestion

NICE (England) report for Infliximab, adalimumab and golimumab for treating moderately to severely active ulcerative colitis after the failure of conventional therapy

The most detailed description of the data used was provided in the HTA report from the English HTA body, NICE, which listed peer-reviewed publications, European public

assessment reports and manufacturer's submissions. The authors searched for clinical effectiveness and cost-effectiveness. Data gaps identified in this report included:

- Available data on hospitalisation outcomes were very limited
- Data on surgical intervention were very sparse, and no trials reported whether surgical outcomes were elective or emergency in nature
- Data on patients with indications that were excluded from the original clinical trials (e.g., patients with ulcerative proctitis, patients with fulminant/acute severe disease, pregnant or lactating women).

Real-world data and Infliximab after the introduction of biosimilars

Infliximab provides an interesting case study for the use of real-world data after the approval of biosimilars for the drug. RCT studies were conducted for the biosimilars, such as, PLANETAS and PLANETRA for ankylosing spondylitis and rheumatoid arthritis, respectively (Yoo *et al.*, 2016; Park *et al.*, 2017). Though none were performed specifically for UC and CD, approval was extended for these indications based on the findings that the pharmacokinetics, efficacy, safety and immunogenicity profiles for the biosimilars were sufficient for the AS and RA indications for Infliximab.

In addition, the introduction of "cheaper" alternatives to Remicade gave rise to a number of budget impact studies in various countries and to a series of studies using real-world data to test efficacy and safety in long-term use. The studies were mostly undertaken for Rheumatoid Arthritis patients rather than IBD patients. However, a prospective observational study was recently published for IBD patients treated with biosimilars (Fiorino *et al.*, 2017). The table below presents the publications and the therapeutic indication and types of data used for each study.

TABLE 6: REAL-WORLD STUDIES OF REMICADE (INFLIXIMAB) AND DATA USED

Study	Therapeutic indication*	Data used
Van Vollenhoven RF, Brannemark S, Klareskog L. Dose escalation of infliximab in clinical practice: improvements seen may be explained by a regression-like effect. Ann Rheum Dis 2004;63:426-430.	RA	The STURE database (Stockholm TNFalpha follow up registry), which collects efficacy and safety data for all patients starting biological treatments at major hospitals in Stockholm, as part of the nationwide registry of antirheumatic therapies in Sweden (ARTIS)
Hokroyd C, Parker L, et al. Abstract	RA	University Hospital Southampton NHS Foundation Trust data from a Rheumatology unit
Waller J, Sullivan E., et al. Assessing physician and patient acceptance of infliximab biosimilars in rheumatoid arthritis, ankylosing spondyloarthritis and psoriatic arthritis across Germany. Patient Preference and Adherence. 3 March 2017	RA, AS, PA	Data from the Adelphi Real World Biosimilars Programme, a real-world cross-sectional study undertaken with German rheumatologists and patients with RA, AS and PA in 2015-2016.
Baraliakos X, Helmann F, et al. Long-term efficiency of infliximab in patients with ankylosing spondylitis: real life data confirm the potential for dose reduction. RMD Open. 2016;2:e000272.	AS	European AS infliximab cohort (EASIC) study data, a follow-up study from the ASSERT trial.

Study	Therapeutic indication*	Data used
Chow A, Khraishi MM, et al. Real-World effectiveness of Infliximab in improving routine assessment of patient index data 3 outcomes: The Canadian experience. 2012 ACR/ARHP Annual Meeting Abstract.	RA	Routine assessment of patient index data e (RAPID3), a pooled index of 3 patient-reported outcomes (PROs).
Preda C, Fulger L et al. Adalimumab and Infliximab in Chrohn's disease – real life data from a national retrospective cohort study. Current Health Sciences Journal. 2016;42(2):117-124.	CD	Data from the archive of the National Insurance Agency in Romania.
Cost-effectiveness of real-world infliximab use in patients with rheumatoid arthritis in Sweden	RA	The STURE database (Stockholm TNFalpha follow up registry), which collects efficacy and safety data for all patients starting biological treatments at major hospitals in Stockholm. Patient-level data on infliximab use were implemented in a Markov cohort model.
Sandborn WJ1, Sakuraba A2, et al. Comparison of real-world outcomes of adalimumab and infliximab for patients with ulcerative colitis in the United States. Curr Med Res Opin 2016;32(7):1233-42.	UC	Medical charts of patient with UC, abstracted by treating physicians in 2014.
O'Donnell S1, Murphy S, et al. Safety of infliximab in 10 years of clinical practice. Eur J Gastroenterol Hepatol. 2011 Jul;23(7):603-6.	UC and CD	Hospital pharmacy records in a single centre.
Tursi A1, Elisei W, et al. Safety and effectiveness of infliximab for inflammatory bowel diseases in clinical practice. Eur Rev Med Pharmacol Sci. 2010 Jan;14(1):47-55.	UC and CD	Hospital data (not specified but seems to be medical charts) in three primary care hospital centres in Bari, Italy.
Fernández-Salazar L, Barrio J, et al. Infliximab use in ulcerative colitis from 2003 to 2013: Clinical practice, safety and efficacy Poster presentation, 2014.	UC	Multicentric and retrospective study which collects clinical data from UC patients treated with IFX in four Spanish hospitals from June 2003 to September 2013

*Crohn's disease (CD), ulcerative colitis (UC), rheumatoid arthritis (RA), ankylosing spondylitis (AP), psoriasis, and psoriatic arthritis (PA)

Data gaps: mapping of available data sources to regulators' and HTA bodies' data needs

Based on the analysis of regulatory and HTA reports, the available data and interviews and a workshop conducted with stakeholders from the regulatory, payer and HTA categories, RCTs have been the preferred data source in the approval process to demonstrate efficacy and safety. However, real-world data has been used to measure long-term safety and efficacy as well as outcomes for various indications for Infliximab. The mapping exercise below pairs real-world data sources listed in the appendix with the data needs identified in the reports for Infliximab.

From the identification of data sources in seven European countries, the list of publications and abstracts provided above where real-world data has been used, the most likely sources of real-world data to meet regulators' and HTA bodies' data needs would be disease registries and clinical records, possibly in electronic form. With some delay, patient-linked administrative data, or administrative claims data, including at least hospital discharge records, medication purchases and mortality registers, could also be used to study efficacy, safety, and long-term effects for Infliximab.

TABLE 7: MAPPING AVAILABLE DATA SOURCES TO DATA NEEDS FOR INFLIXIMAB

Data need	How routinely collected administrative data could be used	How disease registry data could be used	How medical charts, EMR/EHR Records data could be used
Available data on hospitalisation outcomes were very limited	By linking patient-level hospital discharge data with medication purchases and mortality registers, observational studies measuring outcomes for patients receiving infliximab could be conducted.	In a manner similar to the use of the STURE database for Rheumatoid Arthritis, observational studies could be conducted. The STURE database (Stockholm TNFalpha follow up registry), collects efficacy and safety data for all patients starting biological treatments at major hospitals in Stockholm,	Observational studies from participating hospitals or health systems with EHRs could be conducted
Data on surgical intervention were very sparse, and no trials reported whether surgical outcomes were elective or emergency in nature	By linking patient-level hospital discharge data with medication purchases, mortality registers and emergency department data, observational studies measuring outcomes for patients receiving infliximab could be conducted.	Such studies would depend on the level of detail provided by the registry.	Observational studies from participating hospitals or health systems with EHRs could be conducted, provided details regarding emergency vs. elective surgery are available.
Data on patients with indications that were excluded from the original clinical trials (e.g., patients with ulcerative proctitis, patients with fulminant/acute severe disease, pregnant or lactating women).	By linking patient-level hospital discharge data with medication purchases, mortality registers and emergency department data, observational studies measuring outcomes for patients receiving infliximab could be conducted.	The level of detail required to conduct such studies using registry data does not appear to be sufficient. Issues of timeliness are also difficult to overcome registry data is often quite delayed.	Observational studies from participating hospitals or health systems with EHRs could be conducted, provided details regarding concurring diagnoses are available and reliable.

Case study III

Lung cancer

Potential value of big data for the approval of crizotinib

Case study III
Lung cancer
Potential value of big data for the approval of crizotinib

Crizotinib is a treatment for non-small cell lung cancer (NSCLC). Crizotinib is an oral receptor tyrosine kinase inhibitor that blocks enzymes called anaplastic lymphoma kinase (ALK). Crizotinib is only effective in cancer cells that have an overactive version of ALK. Crizotinib has one asymmetric centre of R configuration. It is considered as a class IV compound as per the BCS classification (low permeability, low solubility substance). It is also used when the NSCLC is 'ROS1-positive'. This means that the cancer cells contain changes affecting the gene responsible for the protein ROS1 (EMA 2012).
Crizotinib is the international non-proprietary name and the brand name is Xalkori. The European Commission granted a marketing authorisation valid throughout the European Union for Xalkori on 23 October 2012. The Agency's Committee for Medicinal Products for Human Use (CHMP) concluded that Xalkori treatment has a positive effect on the length of time the patients with ALK -positive NSCLC lived without the disease getting worse, irrespective of whether they were previously treated (EMA 2012).

Data sources

The lung cancer specific data sources identified are mostly registries at either a national or local level. Italy has 38 lung cancer specific data sources including regional assessment report. In the United Kingdom, in addition to registries, a survey and research study were also identified. The following table summarises the data sources that were identified for lung cancer.

TABLE 5: LUNG CANCER DATA SOURCES

	General	LC-specific data source	Name of LC-specific data source
Finland	10	1	The Finnish Cancer Registry
France	11	1	French Cancer Registry
Hungary	8	1	Hungary Cancer Registry
Italy	27	38	38 cancer population registries (at municipal, provincial or regional level)
Norway	6	1	National quality register for lung cancer
Sweden	6	1	National quality register for lung cancer
United Kingdom	23	7	National Cancer Registration and Analysis Service (PHE), Cancer Outcomes and Services Dataset (COSD) Cancer incidence and survival statistics (ONS, ISD Scotland, WCISU, N Ireland Cancer Registry) National Cancer Patient Experience Survey (NHS England) Systemic Anti-Cancer Therapy dataset (SACT) dataset (PHE) National lung cancer audit National Radiotherapy Dataset

Regulatory and HTA data needs

Both national and European assessment reports of Crizotinib were analysed to identify which data was included and what sources were being utilised. Most national and European assessment reports are based on randomised clinical trials and cost-effectiveness studies. The use of real-world data is limited in the assessment reports and seem not to have a crucial role in market authorisation at the timing of the introduction of Crizotinib.

TABLE 6: ASSESSMENT REPORTS FOR CRIZOTINIB

	Assessment Report of Crizotinib
EUROPE	CHMP assessment report of XALKORI (2012) International non-proprietary name: Crizotinib
Finland	Excluded
France	HAS transparency committee Opinion (2013)
Hungary	Excluded
Italy	CHMP assessment report of XALKORI (2012) International non-proprietary name: Crizotinib Multiple Regional Assessment reports
Norway	Excluded
Sweden	The Dental and Pharmaceutical Benefits Agency (TLV 2016)
England	NICE Technology appraisal guidance [TA406]

List of outcome variables included in the data of the assessment reports across the selected countries:

- Overall survival
- Progression-free survival
- Quality of life (EQ-5D)
- Disease control rate
- ICER / Cost per QALY

The CHMP assessment report of Xalkori analyses the clinical efficacy based on available data from a 2 single-arm/uncontrolled multicentre, multinational, open-label, ongoing phase I-II studies (study 1001 and 1005) of crizotinib in patients with advanced (locally advanced or metastatic) ALK-positive NSCLC (EMA 2012).
A consistent high ORR and PFS has been observed with crizotinib in the 2 phase I/II uncontrolled studies and are supported by the preliminary Top-line results of the phase III comparative study (1007).
Example of the clinical data included in in the assessment reports analysed:

- The objective response rate according to therapeutic response in solid tumours (RECIST3) evaluation criteria
- Duration of response to treatment
- Delay in tumour response
- The disease control rate after 8 and 16 weeks of treatment
- Safety
- Objective response rate
- Adverse events

Data gaps: mapping of available data sources to regulators' and HTA bodies' data needs

The analysis of the assessment reports suggest that real-world data source cannot substitute RCTs in the approval process to demonstrate efficacy and safety. However, the value of real-world data seems to be apparent once the treatment has been approved to analyse long-term safety and efficacy for various indications for Crizotinib. The data missing in the assessment report at the time of approval for Crizotinib illustrate some of the areas where real-world data can contribute.

In terms of clinical efficacy, the following data was missing in the EMA report (2012):

- Outcome of post-progression treatments and baseline data of demographics of the patients treated with Crizotinib.
- More thorough analysis of comparators and the difference in overall survival for Crizotinib.
- Additional analysis based on the patient histology.

The EMA report (2012) identifies that there was vital information missing in respect of the clinical safety of Crizotinib, the missing data include:

- Information on patients with hepatic impairment.
- Information on patients with renal impairment.
- Information on elderly patients
- Information on paediatric patients
- Pregnant and lactating women and women of childbearing potential
- Drug interaction with CYP3A inhibitors, inducers, substrates, proton pump inhibitors or H2 antagonists.
- Patients undergoing long-term treatment

Case study IV

Multimorbid patients

Potential value of big data for the approval of medicines in chronic conditions

Case study IV
Multimorbid patients
Potential value of big data for the approval of medicines in chronic conditions

Multimorbidity affects one quarter of the population, and more than half of those over age 65 (Barnett *et al.*, 2012), therefore playing a potentially important role in most patient groups that are seen by health care professionals and that new medicines are developed for. Indeed, 55% of patients with one condition also have at least one other (Barnett *et al.*, 2012), making them fall under the commonly used definition of multimorbid patients (at least two chronic conditions).
As multimorbidities are by definition not a condition *per se*, the case study approach was modified to assess how available data sources can be used to support regulatory decision-making in this area. Selection of conditions was based on clusters of multiple chronic conditions identified in the literature. The body of literature on multimorbidity patterns is growing, with dozens of studies being published over the 20 years. However, some of these only provide associations of chronic conditions, which could randomly occur and therefore not be of relevance to informing decision-making in these diseases. We therefore relied on a systematic review that applied rigorous methodological inclusion and exclusion criteria (Prados-Torres *et al.*, 2014). The systematic review categorised associations of chronic conditions found in included studies into three distinct patterns:

- Cardiovascular and metabolic diseases, with diabetes, hypertension, various forms of heart disease, hyperlipidaemia, and obesity the most common individual components
- Mental health problems, with depression and anxiety the most common individual components
- Musculoskeletal disorders, with arthropathy, back/neck pain, and osteoporosis the most common individual components

For each of the three patterns of multimorbidities, recent marketing authorisation applications for new medicines at the European Medicines Agency (EMA) were reviewed to obtain information on whether multimorbidities were taken into consideration in the clinical evidence development programme or during the evaluation of this evidence, as presented in the publicly available European Public Assessment Reports (EPARs).
Potential multimorbidities in the three patterns include a large number of individual diseases. To make the task of extracting information about multimorbidity considerations manageable, a single medicine was selected for each pattern. The most recently approved novel agent (i.e. excluding generics) for any condition included in one of the three patterns was selected. The three drugs for which information from EPARs was extracted are the following:

- Cardiovascular and metabolic diseases pattern: Suliqua (insulin glargine / lixisenatide) for the treatment of adults with type 2 diabetes mellitus to improve glycaemic control (in combination with metformin)
- Mental health problems pattern: Brintellix (vortioxetine) for treatment of major depressive disorder in adults
- Musculoskeletal disorders pattern: Kevzara (sarilumab) for the treatment of moderately to severely active rheumatoid arthritis (RA) in adult patients who have responded inadequately to, or who are intolerant to one or more disease modifying anti rheumatic drugs (DMARDs) (in combination with methotrexate)

Data sources

We identified 51 data sets in the seven included European countries that are of potential use to study multimorbid patients. Identified data sets represent a variety of different sources and intended uses, such as primary care data sets, routine data bases of in- and outpatient hospital care, death registers, claims data, prescription drugs data (inpatient and outpatient

setting), and more specialised data sets, such as pharmacovigilance data or long-term care data. Data sets were primarily at the national level, although several only have regional coverage, particularly in Italy and the UK. Table 8 breaks down the data sets by country in selected categories.

TABLE 7: OVERVIEW OF IDENTIFIED DATA SETS IN SEVEN EUROPEAN COUNTRIES

	Finland	France	Hungary	Italy	Norway	Sweden	UK
Total data sets	**5**	**5**	**5**	**13**	**6**	**6**	**11**
Primary care	*1*	*-*	*-*	*1*	*1*	*1*	*7*
Hospital (in/outpatient)	*2*	*4*	*3*	*3*	*2*	*2*	*5*
Prescription drugs	*1*	*1*	*1*	*4*	*1*	*1*	*1*
Mortality	*1*	*-*	*1*	*3*	*1*	*1*	*1*
Other	*-*	*2*	*-*	*2*	*2*	*2*	*2*

NB: data sets could be counted in several categories.

Regulatory data needs and mapping of available data

Open questions in each of the EPARs were identified and mapped onto available data sets.

Open questions in the cardiovascular and metabolism multimorbidity pattern

The case study drug for the cardiovascular and metabolism multimorbidity pattern was Suliqua, an insulin glargine / lixisenatide combination indicated (in combination with metformin) for the treatment of adults with type 2 diabetes mellitus to improve glycaemic control in patients where this has not been achieved with metformin or other agents. The drug was approved by the EMA in November 2016, on the basis of a positive benefit-risk balance provided by evidence from two pivotal randomised controlled trials.

The pivotal trials for this drug excluded patients with some of the diseases in the cardiovascular and metabolism pattern, specifically patients with a recent history of stroke, myocardial infarction, unstable angina, or heart failure requiring hospitalization, as well as patients with inadequately controlled hypertension, haemoglobinopathy or haemolytic anaemia. However, these exclusion criteria were deemed appropriate by the reviewers, who considered the trial populations as representative for the diabetes population benefitting from the drug. While patients with recent cardiovascular events or uncontrolled hypertension were excluded, the trials did include patients with a history of cerebrovascular events or cardiovascular risk factors, addressing the question of efficacy and safety in patients with some of the most prevalent co-existing chronic conditions.

All open questions and required data, as well as an indication of which data bases could be used to answer them, are described below and summarised in Table 9.

Open question 1: efficacy in obese patients

Based on the weight effects of the drug, the review committee stated that ‘The main target population for Suliqua is expected to be patients eligible for initiation or intensification of insulin treatment and where there is a need to avoid (further) weight increase.’ The pivotal trials included obese patients, and a pre-specified meta-analysis of their results stratified by BMI (< and >30) was presented. In fact, the mean BMI of patients in pivotal trials was >30 and there are questions regarding the efficacy in non-obese patients. However, the potentially higher efficacy in obese vs. non-obese patients is not reflected in the indication for which marketing authorisation was granted, despite uncertainties regarding the efficacy in different weight groups were mentioned by the reviewers. While not explicitly stated by the reviewers (and not included in post-authorisation study requirements), an open question derived from reviewer comments is therefore:

- What is the efficacy of Suliqua in non-obese patients?

Data required to study this question include the following:

- Cohort of patients with type 2 diabetes mellitus taking Suliqua

- BMI information
- Outcomes: change in HbA1c from baseline to Week 30 (primary endpoint used in pivotal trials)

Potentially useful data bases to answer this question include the Finnish register of primary health care visits (although it is unclear whether BMI would be available and to what extent lab results are included), the French inter-scheme consumption data (données de consommation inter-régimes [DCIR]) (although lab data availability questionable), as well as a range of UK data sources. The Italian regional hospitalisation databases use links to prescription medicines data sets. Again, availability of lab results is unclear, but these data sets are linkable to others using a patient identifier. This includes local primary care data sets, which could be more likely to hold this information. Finally, an ambulatory care database exists in Italy, which again can be linked to other datasets.

Open question 2: drug-drug interaction

Given that Suliqua is a fixed-dose combination of already approved agents (insulin glargine and lixisenatide), the manufacturer relied on evidence of no drug-drug interaction of its components as monotherapies with other drugs commonly taken by patients with type 2 diabetes. In particular, the influence of lixisenatide monotherapy on ramipril and atorvastatin (among others) was evaluated. However, no additional studies have been conducted to investigate drug-drug interactions of the combination drug Suliqua. While not explicitly mentioned in the assessment report of the EMA, the lack of drug-drug interaction data for the combination drug raises the following as an open question that could be addressed using RWE:

- Are there any safety concerns regarding drug-drug interaction of Suliqua with other medicines commonly taken by patients with type 2 DM?

Data required to answer this include the following:

- Cohort of patients with type 2 diabetes mellitus taking Suliqua and taking other medicines
- Adverse events data

The question requires the combination of adverse events data, likely to be held in hospitalisation data bases or emergency care data sets, and prescription drugs. The Italian regional hospitalisation data can be linked to other data sets, including prescription medicines. This would allow researchers to identify patients taking the drug of interest who were admitted to hospital with an adverse event, with additional information on other medicines taken by these patients. Medication data are held in regional datasets on publicly funded medication purchases, which can be linked to hospitalisations.

Norway and Sweden: patient administrative systems exist that include inpatient and outpatient visits. However, it is unclear to what extent medication data are available from these data sets or whether they can be linked.

Some of the UK primary care data bases include links to emergency care visits and hospitalisations, as well as prescription drug data.

Open question 3: safety in special population with associated disease

In the discussion on clinical safety, a higher rate of adverse events in patients with moderate renal impairment was noted. However, the committee noted that the population with moderate renal impairment was small in the studies, and no patients with severe renal impairment were included in any study. The missing information on use of the drug in patients with severe renal impairment is included under ‘safety concerns’ in the risk management plan. An open question regarding the safety of the drug in a special population with an associated disease is therefore the following:

- Is the use of Suliqua safe in patients with severe renal impairment?

Data required to answers this question include the following:

- Cohort of patients with type 2 DM and severe renal impairment, taking Suliqua (overriding special warning and precautions of use)
- Treatment emergent adverse events data

This is a classical pharmacovigilance question that could be addressed using existing systems that collect data on adverse events, such as the Italian drug registry (maintained by the Italian Medicines Agency, AIFA), as well as other data bases that combine information on drug use with adverse events data (hospital admissions, emergency care).

TABLE 8: DATA NEEDS AND AVAILABLE DATA SOURCES FOR INSULIN GLARGINE / LIXISENATIDE FIXED-DOSE COMBINATION

Open question	Data required	Potential data sets
What is the efficacy of Suliqua in **non-obese patients**?	• Cohort of patients with type 2 diabetes mellitus taking Suliqua • BMI information • Outcomes: change in HbA1c from baseline to Week 30 (primary endpoint used in pivotal trials)	• Finnish register of primary health care visits • French inter-scheme consumption data (données de consommation inter-régimes [DCIR]) • Italian regional hospitalisation databases, linked to prescription and primary care data sets • Norwegian and Swedish patient administrative systems (in- and outpatient) when linked to prescriptions database • English General Practice Research Datalink (CPRD), linked to Hospital Episodes Statistics (HES) • Scottish Primary Care Clinical Informatics Unit (PCCIU) • UK The Health Improvement Network (THIN) • English QResearch database • English regional Consultation in Primary Care Archive (CiPCA) • UK CALIBER database • Welsh SAIL databank • UK Research One/TPP database
Are there any safety concerns regarding **drug-drug interaction** of Suliqua with other medicines commonly taken by patients with type 2 DM?	• Cohort of patients with type 2 diabetes mellitus taking Suliqua and taking other medicines • Adverse events data	• Italian regional hospitalisation databases, linked to prescription data sets • Norwegian and Swedish patient administrative systems (in- and outpatient) when linked to prescriptions database • English General Practice Research Datalink (CPRD), linked to Hospital Episodes Statistics (HES) • UK The Health Improvement Network (THIN) • English QResearch database • UK CALIBER database • Welsh SAIL databank • UK Research One/TPP database
Is the use of Suliqua **safe** in patients with severe renal impairment?	• Cohort of patients with type 2 DM and severe renal impairment, taking Suliqua (overriding special warning and precautions of use)	• Italian drug registry at the Italian Medicines Agency (AIFA) • Italian regional hospitalisation databases, linked to prescription data sets • Norwegian and Swedish patient administrative systems (in- and outpatient) when linked to prescriptions database • English General Practice Research Datalink (CPRD), linked to Hospital Episodes Statistics (HES) • UK The Health Improvement Network (THIN) • English QResearch database

	• Treatment emergent adverse events data	• UK CALIBER database • Welsh SAIL databank • UK Research One/TPP database

Open questions in the mental health multimorbidity pattern

The case study drug for the mental health multimorbidity pattern was vortioxetine, indicated for treatment of major depressive episodes in adults. In the discussion on clinical efficacy in the EPAR on vortioxetine, the committee noted that 'Overall the patient population is considered to be a rather homogen sub-population of the general MDD population since frequently occurring comorbidities and suicidal thoughts/suicidality, amongst others, were excluded. This is considered acceptable to reduce confounders and facilitation of evaluation of the pure antidepressant effect.' While this indicates the committee was content with the restricted nature of the trial population, open questions remain.

Open question 1: withdrawal rate in older patients

The first question was included as an objective for ongoing and planned studies in phase IV and is being addressed through a post-authorisation safety study (non-interventional cohort).

- What is the rate of withdrawal of vortioxetine due to lack of efficacy in patients aged 75 and over?

Required data for this question include:

- Use and discontinuation of vortioxetine in patients with major depressive episodes
- Reasons for discontinuation

Available data sets might not be particularly well suited to answer this open question, since it requires some information on the reason for discontinuing the medicine (e.g. whether this is due to lack of efficacy – which is the question of interest – or due to other factors, such as allergic reactions, administration mode etc.). Such information may be included in the treating physician's notes, therefore datasets that rely on medical records, such as The Health Improvement Network (THIN) database in the UK, could be relevant. However, administrative data sets might not include free text. For example, the Scottish Primary Care Clinical Informatics (PCCIU) database does not contain free text and would therefore be unlikely to be of value to answer this research question.

Another potentially valuable data source is the Finnish register of primary health care visits, which contains information on 'events', i.e. primary care visits, recording the reasons for visits and the outcome of the assessment for need of treatment.

Without explicit notes on the reasons for discontinuation, other means of ascertaining treatment withdrawal due to lack of efficacy could include analyses of symptoms, which, if not improving or deteriorating over time despite receiving treatment, could indicate that the patient is not responding to the treatment and subsequent withdrawal could therefore be associated to it. The English regional Consultation in Primary Care Archive (CiPCA) contains information on symptoms, as well as prescriptions, and uses a patient ID to track patients over time.

Open question 2: efficacy and safety in older patients

- What is the efficacy and safety of vortioxetine dosed at over 10mg/day in patients aged 65 years or older?

Data requirements to answer this question include:

- Cohort of patients aged 65 years or older being treated with vortioxetine >10mg/day for major depressive episodes
- Dosing data of drug
- Outcomes data: MADRS or Hamilton Depression Rating Scale (primary endpoint tools used in pivotal trials)
- Other possible outcomes: cognitive function, health-related quality of life

Due to the specific outcomes required to measure efficacy in the treatment of major depressive episodes (patient-reported outcomes or physician-assessed scales), the most likely data sources of value are those specialising in mental health. The French PMSI PSY database is a candidate (if dosing data of medications and relevant outcomes are included). Another candidate is the English CPRD (containing medication data), which can be linked to

the specialised Mental Health Dataset.
The English regional CiPCA also includes patient-reported outcomes.

Open question 3: safety in older patients with co-medication

In their conclusions on clinical pharmacoepidemiology, the committee noted that 'No clinically relevant pharmacokinetic or pharmacodynamic interactions were observed following co-administration of Vortioxetine with aspirin, warfarin, oral contraceptives, or the CNS-active compounds alcohol, diazepam, or lithium.' However, to address a safety concern for missing information on the use of vortioxetine in patients with comorbid Parkinson's disease, information on potential drug interaction with selegiline and rasagiline is included in the summary of product characteristics.
Furthermore, uncertainty regarding the safety of vortioxetine in patients >75 years was mentioned, as these patients may be additionally affected by hyponatraemia (low sodium level in blood) while receiving co-medication (e.g. diuretics). An open question (not explicitly stated in the EPAR) is therefore the following:

- What is the safety of vortioxetine in patients >75 years who are taking co-medications?

Data required to answer this question include the following:

- Cohort of patients aged >75 years being treated with vortioxetine for major depressive episodes while also receiving other medication (e.g. diuretics)
- Lab data: sodium levels

This open question addresses a specific safety concern (dropping sodium levels in the blood). Relevant data sources therefore need to include laboratory data, as well as information on medication. Some of the primary care data sets in the UK include both test results and medication data, including the PCCIU, THIN and QResearch databases, as well as the more comprehensive Welsh SAIL databank, which also includes secondary care data.

Open question 4: efficacy and safety in special population with associated disease

Under safety concerns, the applicant noted that information was missing on safety for use in patients aged 75+ years, patients with a history of mania or hypomania, patients with severe renal or hepatic impairment, and in patients with comorbid Alzheimer's disease, Parkinson's disease, multiple sclerosis, and stroke. The committee included in the discussion on the uncertainty in the knowledge about unfavourable effects that 'Patients with a history of mania or hypomania were excluded from the studies; therefore, the possible switch from depression to mania as a result of Vortioxetine treatment has not been investigated. A PASS will provide further information on the use, efficacy, adverse events and withdrawals in patients with (a history of) mania/hypomania.' Note that the risk management plan only included the study of the use of the drug in this group, but not its efficacy or adverse events in this population.
Open questions regarding the efficacy, safety and withdrawal rates therefore include the following:

- What is the efficacy, rate of adverse events, and rate of withdrawals in patients with (a history of) mania/hypomania?

Data needs to answer this question include the following:

- Vortioxetine use in patients treated for major depressive episodes who have a history of mania/hypomania
- Outcomes data: MADRS or Hamilton Depression Rating Scale (primary endpoint tools used in pivotal trials)
- Other possible outcomes: cognitive function, health-related quality of life
- Adverse events data

This is a question with complex data requirements, as it involves outcomes data that are not typically included in routine databases (and instead require specialist data sets) as well as

adverse events data that would typically come from hospital data sets with information on emergency visits and admissions. One of the data sets identified in our search could fulfil all requirements through data linkage: the English General Practice Research Datalink (CPRD) can be linked to the Metal Health Dataset (MHDS) and the Hospital Episode Statistics (HES).

TABLE 9: DATA NEEDS AND AVAILABLE DATA SOURCES FOR VORTIOXETINE

Open question	Data required	Potential data sets
What is the rate of **withdrawal of vortioxetine** due to lack of efficacy in patients **aged 75 and over**?	• Use and discontinuation of vortioxetine in patients with major depressive episodes • Reasons for discontinuation	• Finnish register of primary health care visits • UK The Health Improvement Network (THIN) database • English regional Consultation in Primary Care Archive (CiPCA)
What is the efficacy and safety of vortioxetine dosed at over 10mg/day in patients **aged 65 years or older**?	• Cohort of patients aged 65 years or older being treated with vortioxetine >10mg/day for major depressive episodes • Dosing data of drug • Outcomes data: MADRS or Hamilton Depression Rating Scale (primary endpoint tools used in pivotal trials) • Other possible outcomes: cognitive function, health-related quality of life	• French PMSI PSY (psychiatry-specific database) • English General Practice Research Datalink (CPRD), linked to the Metal Health Dataset (MHDS) • English regional Consultation in Primary Care Archive (CiPCA)
What is the safety of vortioxetine in **patients >75 years who are taking co-medications**?	• Cohort of patients aged >75 years being treated with vortioxetine for major depressive episodes while also receiving other medication (e.g. diuretics) • Lab data: sodium levels (hyponotraemia is low sodium level in blood)	• Scottish Primary Care Clinical Informatics Unit (PCCIU) • UK The Health Improvement Network (THIN) • English QResearch database • Welsh SAIL databank
What is the efficacy, rate of adverse events, and rate of withdrawals in **patients with (a history of) mania/hypomania**?	• Vortioxetine use in patients treated for major depressive episodes who have a history of mania/hypomania	• English General Practice Research Datalink (CPRD), linked to the Metal Health Dataset (MHDS) and Hospital Episode Statistics (HES)

	• Outcomes data: MADRS or Hamilton Depression Rating Scale (primary endpoint tools used in pivotal trials) • Other possible outcomes: cognitive function, health-related quality of life • Adverse events data	

Open questions in the musculoskeletal disease multimorbidity pattern

For the musculoskeletal disease multimorbidity pattern, the case study drug was Kevzara (sarilumab), which is indicated in combination with methotrexate for the treatment of moderately to severely active rheumatoid arthritis (RA) in adult patients who have responded inadequately to, or who are intolerant to one or more disease modifying anti rheumatic drugs (DMARDs).

Open question 1: efficacy in obese patients

Subgroup analyses showed reduced efficacy in patients with a BMI =>30. There is therefore a need to answer the following question:

- What is the efficacy of sarilumab as add-on or monotherapy in patients with BMI=>30?

Required data for this question include:

- Cohort of patients with BMI=>30 being treated with sarilumab for moderately to severely active rheumatoid arthritis
- Information on concomitant medication
- Outcomes data: American College of Radiology 20% improvement score (ACR20); Health Assessment Questionnaire-Disability Index (HAQ-DI); modified Total Sharp Score (mTSS); Disease Activity Score 28 Joints using Erythrocyte Sedimentation Rate (DAS28-ESR) (primary endpoint tools used in pivotal trials)

Outcomes used in pivotal trials are highly disease-specific and unlikely to be included in primary care data or hospital data sets. No rheumatology-specific data sets were included in our sample of 51 databases. We could not identify data sources from this set that could be used to answer the open question regarding efficacy of sailumab in obese patients for the endpoints used in the pivotal trials.

Open question 2: efficacy and safety in elderly patients

The EPAR mentioned missing information on the drug's safety in the elderly. The number of elderly patients in the pivotal trials was also reported, but the risk minimisation activity relating to this was restricted to 'appropriate SmPC [Summary of Product Characteristics] statements/information'. Post-authorisation study requirements were not specified, but the lack of information on efficacy and safety in older patients raises the following question:

- What is the efficacy and safety of sarilumab in elderly patients?

Required data for this question include:

- Cohort of elderly patients aged 65 years or older being treated with sarilumab for moderately to severely active rheumatoid arthritis
- Outcomes data: American College of Radiology 20% improvement score (ACR20); Health Assessment Questionnaire-Disability Index (HAQ-DI); modified Total Sharp Score (mTSS); Disease Activity Score 28 Joints using Erythrocyte Sedimentation Rate (DAS28-ESR) (primary endpoint tools used in pivotal trials)

Again, the outcomes used in regulatory trials are disease-specific and unlikely to be included in primary care or hospital data sets (see open question 1 above).

Open question 3: safety in subpopulation with associated disease

Given the increased risk of the rheumatoid arthritis population of cardiovascular disease, the potential risk of cardiovascular events was highlighted as important: 'Similarly, a relationship between lipid increase and CV risk during sarilumab treatment cannot be ruled out at present, considering that the RA population is at higher risk of CV diseases compared to the general population. Although not resolvable at present, this risk may be taken into account by including a warning on the increased risk of cardiovascular disorders in patients with RA in section 4.4 of the SmPC.' These considerations are also at the heart of the only post-authorisation study for this drug. The study's focus is on safety events, including serious infections, lipid abnormalities and increased risk of major cardiovascular events, gastrointestinal perforations, malignancy, as well as the use of sarilumab in pregnant

women. The open question is the following:

- What is the risk of patients taking sarilumab experiencing lipid abnormalities and major cardiovascular events?

Required data for this question include:

- Cohort of patients being treated with sarilumab for moderately to severely active rheumatoid arthritis
- Outcomes data: lipid levels; cardiovascular events

This open question could be addressed through data sets that include medication data, diagnoses, and two different types of outcomes: lab results (for lipid levels) and emergency admissions (for cardiovascular events).
The Italian database on publicly funded medication purchases could be linked to emergency services data to identify cardiovascular events. Hospital admissions (regional datasets) can also be linked to medication data, allowing a relevant cohort to be identified.
UK databases that could be used to address this question include, most importantly, CALIBER, a national database that links primary and secondary care data with prescription drug data and a cardiovascular disease-focused registry (MINAP). Other relevant data sets in the UK include CPRD, THIN, Research One, the Welsh SAIL databank, the Scottish PCCIU (all including test results and emergency admissions), and QResearch (test results only).
Swedish and Norwegian inpatient administrative systems (claims data) could be of value if linked to prescription data.

TABLE 10: DATA NEEDS AND AVAILABLE DATA SOURCES FOR SARILUMAB

Open question	Data required	Potential data sets
What is the efficacy of sarilumab as add-on or monotherapy in **patients with BMI=>30**?	• Cohort of patients with BMI=>30 being treated with sarilumab for moderately to severely active rheumatoid arthritis • Information on concomitant medication • Outcomes data: American College of Radiology 20% improvement score (ACR20); Health Assessment Questionnaire-Disability Index (HAQ-DI); modified Total Sharp Score (mTSS); Disease Activity Score 28 Joints using Erythrocyte Sedimentation Rate (DAS28-ESR) (primary endpoint tools used in pivotal trials)	None identified, as outcomes used in pivotal trials are highly disease-specific and unlikely to be included in the data sets identified.
What is the efficacy and safety of sarilumab in **elderly patients**?	• Cohort of elderly patients aged 65 years or older being treated with sarilumab for moderately to severely active rheumatoid arthritis • Outcomes data: American College of Radiology 20% improvement score (ACR20); Health Assessment Questionnaire-Disability Index (HAQ-DI); modified Total Sharp Score (mTSS); Disease Activity Score 28 Joints using Erythrocyte Sedimentation Rate (DAS28-ESR) (primary endpoint tools used in pivotal trials)	None identified, as outcomes used in pivotal trials are highly disease-specific and unlikely to be included in the data sets identified.
What is the **risk** of patients taking sarilumab experiencing lipid abnormalities and	• Cohort of patients being treated with sarilumab for	• Italian database on publicly funded medication purchases (linked to medication data)

major **cardiovascular** events?	moderately to severely active rheumatoid arthritis • Outcomes data: lipid levels; cardiovascular events	• Italian regional hospital datasets (linked to medication data) • UK CALIBER database (linked to CPRD, HES, and MINAP) • English General Practice Research Datalink (CPRD), linked to the Hospital Episode Statistics (HES) • Scottish Primary Care Clinical Informatics Unit (PCCIU) • UK The Health Improvement Network (THIN) • Welsh SAIL databank • UK Research One/TPP database • English QResearch database • Norwegian inpatient administrative systems data • Swedish inpatient administrative systems data

4.1 General (non-disease specific) data sources

FINLAND					
Care Register for Health Care (inpatient and day hospital)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (up to 2017)	**Data on the patient/client:** • Personal identity number • Municipality of residence • Country code for non-Finnish residents **Data on start of care:** • Referring party • Code and code extension of the institution that referred the patient • Waiting list entry date and date of admission • Type of admission • Route of admission • Code of the place of	**Data on the treatment received by the patient/client and on the grounds for a client relationship:** • Reason for seeking care • Diagnoses (ICD10-CM codes) • External cause • Type of accident • Need for care on date of admission/discharge/count • Procedures and interventions • Decision on long-term care (yes/no) • Patient has an advanced cardiac condition (yes/no) • Patient is a psychiatric patient (yes/no) • Number of home days **Data on discharge from care:** • Date of discharge • Further treatment/which services	• Length of stay (day) • Hospitalization type (ordinary admission vs day surgery) • Costs	• Hospitalizations • Hospitalization costs • Intra-hospital mortality • Length of stay	Hospitalisation data (individual level).

	discharge (if the person was admitted from another institution)	• Code and code extension of the institution of further treatment			
Prescribed medication					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (2005-2017)	• Patient ID • ATC code • Number of units • Date purchased • Cost	• ATC code	• Medication purchases, indications of comorbidities	• Resource use	
Care Register for Health Care (outpatient)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (up to 2017)	• Patient ID • Number of services/visits • Date provided • Cost	• Finnish Procedure codes	• Cost	-	Outpatient hospital visits (individual level data)

Register of primary health care visits					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (since 2011 all care in Finland)	• Client's personal identity number (encrypted in the register) • Client's municipality of residence • Postcode of the client's place of residence • Service provider	**Information on 'service event':** • service provider in the service event • client • contact by client • assessment of the need for treatment • appointment data: date and time of appointment booking, date and time of booked appointment, occupation, service type, contact type • data on the content of the service event, such as reason for visit, procedures, and follow-up care • Event record code	• Procedures and interventions	• Data on the content of the service event, such as reason for visit, procedures, and follow-up care	

	• Service unit of the service provider	**Data on contact type and assessment of the need for treatment:** • Date and time of contact • Date and time of assessment of the need for treatment • Occupation (person assessing the need for treatment) • Urgency of care • Type of visit • Outcome of assessment of the need for treatment **Service event data:** • Date and time, service event starting time • Date and time, service event closing time • Occupation (person delivering care) • Service type • Contact type • Client group • Urgency • Type of visit • First visit • Reasons for visit/diagnoses • External cause (accidents) • Type of accident • Procedures and interventions • Vaccination data and medication data • Dental health care, DMFT index • Dental health care, CPI index **Follow-up care:** • Service event cancellation • Reason for service event cancellation			
Mortality registry					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**

National (2005-2017)	• Age • Gender	• Cause of death	-	• Cause of death • Deaths/death rate	
Hospital emergency care					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Local	**Individual. Not available outside hospitals.**				
Long term care/ nursing care in institution					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National. 2005 - 2017	**Individual**				
Assistance services and social services					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
2006 - 2013					
Personal care services at home					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Capital city areas only. 2006 – 2013					
Special reimbursement register					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National		Encompasses 44 specific chronic diseases or conditions			

FRANCE

Programme for the medicalization of information systems (programme de médicalisation des systèmes d'information [PMSI]) of the Système national d'information inter-régime de l'assurance maladie (SNIIR-AM)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National, 86% of insured patients until 2009, 100% after 2009 (coverage: 10 years plus current year)	• Individual code • Age • Gender • Place of residence: municipality code • Universal Health Coverage • Affiliated insurance scheme	• Diagnoses (ICD10-CM codes) • Discharge status • DRG • Procedures • Hospital and unit identifier • score of severity for stays in intensive care units (Simplified Acute Physiology Score II)	• Length of stay (day) • Hospitalization type (ordinary admission vs day surgery) • Costs	• Hospitalizations • Intra-hospital mortality • Length of stay • Date of death (linked since 2009) • Mortality rates (linked since 2009)	Mortality data linked since 2009 No access to the individual code, making it difficult to match patients to hospital cohorts

PMSI MCO for médecine chirurgie obstétrique

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National, separate database for these areas (coverage: 10 years plus current year)	• Individual code	• Medicine, Surgery and Obstetrics hospitalisations	-	-	

Inter-scheme consumption data (données de consommation inter-régimes [DCIR])

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	• Age • Gender • Place of residency	• ATC codes (unconfirmed) • long-term diseases (LTD)	• Outpatient visits (?)	• Outpatient visits (?) • Resource use	The DCIR contains several data sets, including detailed

	• Insurance scheme • Benefit from the Universal Health Coverage	• Occupational accidents, sick leave and occupational diseases	• Name, form, quantity of prescribed drugs		demographic information, data on prescriptions, outpatient visits and long-term care
PMSI PSY for psychiatrie					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (10 years plus current year)	• Individual code	-	-	-	
PMSI HAD for hospitalisation à domicile					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (10 years plus current year)	• Individual code	-	-	-	Database for home care
Inter-scheme consumption data (données de consommation inter-régimes [DCIR]) - demographics					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National	Age, gender, place of residency, insurance scheme, benefit from the Universal Health Coverage				
Inter-scheme consumption data (données de consommation inter-régimes [DCIR]) – prescribed medication					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National. Individual.					

Inter-scheme consumption data (données de consommation inter-régimes [DCIR]) - outpatient					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Individual	Age Gender Residency Insurance affiliation scheme				
Système national d'information inter-régime de l'assurance maladie (SNIIR-AM) - mortality					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National. Individual. Linked with date of death since 2009 Depth 3 years plus current year		Date of death		Date of death, mortality rates	
Inter-scheme consumption data (données de consommation inter-régimes [DCIR]) – long term care, home care					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
long-term diseases (LTD), occupational accidents, sick leave and occupational diseases are also included in this database, at the individual	Age Gender Residency Insurance affiliation scheme				

level					
PMSI SSR for soins de suite et de réadaptation – rehabilitation					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Individual. 10 years plus current year.					

HUNGARY

Hospital discharge register (HDR), National Health Insurance Fund Administration of Hungary					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (2005-2014)	• Age • Gender • residency	• Diagnoses (ICD10-CM codes) • Discharge status	• Length of stay (day) • Hospitalization type (ordinary admission vs day surgery)	• Hospitalizations • Intra-hospital mortality • Length of stay	
Prescribed medicine, with costs, National Health Insurance Fund Administration of Hungary					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (2005-2014)	• Patient ID	• ATC code	• Number of units • Date purchased • Cost	• Resource use	
Hospital emergency care: National Health Insurance Fund Administration of Hungary					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (2008 onwards	• Age • Gender • Residency	-	-	-	No further information available

Outpatient care in specialized health care institutions, National Health Insurance Fund Administration					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (2005-2014)	• Age • Gender • Residency	-	-	-	No further information available
Date of death statistics					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (2005-2015)	-	• Date of death (cause of death not available, although it is collected)	-	• Date of death • Mortality rates	
Care register for social welfare					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
2005 - 2014					
Files for social care and primary health care Register of Primary Health Care Visits					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Local					
Home special nursing and therapy service					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
					There are statistics provided on home special nursing and therapy service on the Hungarian Central Statistical Office website (https://www.ksh.hu/stadat_annual_2_4), but there is no description of the

					name of the database and variables included.

ITALY

Hospital Discharge Database (SDO)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National (2001-2015)	• Age • Age class • Gender • Citizenship • Residence (Region/District/ Municipality) • Marital status • Educational level	• Diagnoses (ICD9-CM codes; ICD9-CM category for main diagnosis) • Procedures (ICD9-CM procedure codes)	• DRG code • DRG type (medical or surgical) • Length of stay (day) • Hospitalization type (ordinary admission vs day hospital) • Payer (e.g. NHS, patient, other)	• Hospitalizations • Hospitalization costs • Intra-hospital mortality • Length of stay	

Hospitalisation (regional databases)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
Regional (all)	• Patient ID • Gender • Age • Citizenship • Residence (Region/District/ Municipality) • Marital status • Educational level • (in some areas) Employment status • (in some areas) Job type	• Diagnoses (ICD9-CM codes; ICD9-CM category for main diagnosis) • Procedures (ICD9-CM procedure codes)	• Hospitalization type (ordinary admission vs day hospital) • Time in ED • Resulting hospital admission (yes/no) • Payer (e.g. NHS, patient, other) • Re-hospitalization (using ID codes) • Linkable to other databases	• Hospitalizations • Hospitalization costs • Intra-hospital mortality • Length of stay • For Emergency care patients: Triage code and exit code allow linking the record to a resulting hospital admission • Length of time in Emergency Department • Through patient ID linked to medications, mortality, and	Regional databases that can be linked using the patient ID: • emergency services • nursing care and hospice • home health care • primary care services (limited, and dictated by local level practices) • disabilities • CEDAP – pregnancy and birth database • patient satisfaction surveys

				ambulatory care databases: ○ Mortality ○ Comorbidities ○ resource use ○ follow-up care ○ previous care ○ re-hospitalizations ○ long-term outcomes ○ complications of the hospitalized patients	
Piemont regional hospital database					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Regional (Piemont)	• Patient ID • Gender • age • Citizenship • Residence (Region/District/ Municipality) • Marital status • Educational level • Employment status • Job type	• Diagnoses (ICD9-CM codes; ICD9-CM category for main diagnosis) • Procedures (ICD9-CM procedure codes) • Use of ROBOT during procedures • Checklist operating room • Pain assessment • Cancer stage (only for ICD9-CM 140.0-190.9 and 193-199.1)	• Hospitalization type (ordinary admission vs day hospital) • Admission over day 365 days • Payer (e.g. NHS, patient, other) • Re-hospitalization (yes/no) • Cancer drugs (provided during the hospital stay)	Same as above Additional information about cancer patients	

		• Baby feeding type			

Publicly funded medication purchases

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
Regional (2004-2015)	• Patient ID	• ATC code • Indications of comorbidities	• Medication purchases • Number of units • Date purchased • Cost	• Resource use	Prescription drug database. Regional databases that can be linked using the patient ID: • emergency services • nursing care and hospice • home health care • primary care services (limited, and dictated by local level practices) • disabilities • CEDAP – pregnancy and birth database • patient satisfaction surveys

Publicly funded medications administered during hospital stays

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
Regional (2004-2015)	• Patient ID	• ATC code • Indications of comorbidities	• Medications administered in the hospital • Number of pills/vials, etc. • Date purchased • Cost	• Resource use	Drugs administered during hospitalisation. Regional databases that can be linked using the patient ID: • emergency services • nursing care and hospice • home health care

					• primary care services (limited, and dictated by local level practices) • disabilities • CEDAP – pregnancy and birth database • patient satisfaction surveys
Drug Registry (Italian Medicines Agency - AIFA)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National	- (aggregate level data)	- (aggregate level data)	(aggregate level data) There are three observatories: • Pharmacovigilance – collects all data on adverse effects and oversees risk-benefits profiles for all medications. This data is then integrated into the European EUDRA Vigilance database • Medication use – monitors all medications covered by the national health system and transmits data monthly to the regions, using predefined indicators regarding expenditure • Clinical trials – assures the monitoring of all clinical trials conducted in Italy and approved by local Ethics Committees • In addition to publicly-funded medicines, AIFA purchases aggregate out-of-pocket sales data from IMS to monitor total expenditure. They publish the results of their analysis of medication use in Italy every year, available on the website.	(aggregate level data) • Adverse events through pharmacovigilance observatory: collects all data on adverse effects and oversees risk-benefits profiles for all medications. This data is then integrated into the European EUDRA Vigilance database	

Ambulatory Care Database					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Regional (all, 2004-2015)	• Patient ID	• ICD-9 Procedure codes • Number of services/visits • Date provided	• Cost • Specialty visits by ICD-9 procedure code • Number of services/visits	• Resource use • Adhesion to clinical guidelines	Regional databases that can be linked using the patient ID: • emergency services • nursing care and hospice • home health care • primary care services (limited, and dictated by local level practices) • disabilities • CEDAP – pregnancy and birth database • patient satisfaction surveys
National Social Security Institute (INPS)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National	• Age • Gender • Job sector (Agriculture, Industry, Handicraft, Services)	• 20 disease categories (INPS classification) in a separate database (GASAN)	-	• Disability pension • Sick leave • Disability allowance	Individual level data (record linkage with GASAN database through fiscal code)
Exemptions for pathology					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Local (local health authority - ASL)	• Patient ID • Exemption code by condition (Italian system)	• Exemption code (009 based on DM 329/99)	• Excuses patient from co-payments and medication purchases	• Helps identify patients with particular	

				pathologies, identify comorbidities	

Mortality registry (ISS)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National (1980-2003; 2006-2012)	(aggregate level data) • Age class • Region	(aggregate level data) • Cause of death (ICD9-CM codes until 2002; ICD10 thereafter)	-	(aggregate level data) • Cause of death • Deaths/death rate	

Mortality registry (ISTAT)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National (2003-2014)	(aggregate level data) • Age • Age class • Birth year • Gender • Citizenship • Area/Region/ District • Marital status (if married, wedding year and age class of living spouse) • Educational level	(aggregate level data) • Cause of death (European Shortlist for Causes of Death – COD) • Comorbidities (average number) • Multiple causes of death	-	(aggregate level data) • Cause of death • Deaths/death rate	

Regional mortality registries (Veneto, Emilia-Romagna, Tuscany, Abruzzo, Umbria)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
Regional (Veneto, Emilia-Romagna, Tuscany, Abruzzo, Umbria)	(aggregate level data) • Age/age class • Gender	(aggregate level data) • Cause of death (ICD9/ICD10) (Veneto, Umbria) • Cause of death (Tuscany, Abruzzo)	-	(aggregate level data) • Cause of death • Deaths/death rate	

		• None (Emilia-Romagna)			
Local mortality registries (ASL Vercelli and local districts)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Local (ASL Vercelli and local districts)	(aggregate level data) • Gender	• Cause of death; cause of death (ICD9/ICD10)	-	(aggregate level data) • Cause of death • Deaths/death rate	
Other regional databases that can be linked using patient ID (see comments column)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
	Patient ID				· emergency services · nursing care and hospice · home health care · primary care services (limited, and dictated by local level practices) · disabilities · CEDAP – pregnancy and birth database · patient satisfaction surveys

NORWAY

Inpatient registry (Norwegian patient registry)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	• Age • Gender • Patient's region • Patient's municipality	• Diagnoses (ICD-10) • Surgical procedure codes • Medicinal procedure codes	• DRG • Bed days	• Hospitalizations	Registries at the National board of health and welfare

Norwegian Prescription Database (NorPD)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	• Age • Gender	-	• Prescriptions administered at outpatient pharmacies	-	Registries at the National board of health and welfare

Causes of death registry (Dødsårsaksregisteret)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	• Age • Gender • Location at death • Country of origin	• Diagnoses (ICD-10)	-	• Date of death • Cause of death	Registries at the National board of health and welfare

Statistics Norway

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	• Age • Gender • Educational level • Income • Country of birth	-	-	-	

Social Insurance Agency					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National	• Age • Gender	-	• Disability and sick leave payments	• Disability pension • Sick leave	
Patient administrative systems					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Regional	• Age • Gender	• Diagnoses (ICD-10) • Procedures	• Inpatient care • Outpatient visits • Primary care visits	• Hospitalization or visits	

SWEDEN

Inpatient registry

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	• Age • Gender	• Diagnoses (ICD-10) • Procedures (KKÅ97)	• DRG • Bed days	• Hospitalizations	Registries at the National board of health and welfare

Prescription registry

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	• Age • Gender	-	• Prescriptions administered at outpatient pharmacies	-	Registries at the National board of health and welfare

Causes of death registry

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	• Age • Gender	-	-	• Date of death • Cause of death	Registries at the National board of health and welfare

Statistics Sweden

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	• Age • Gender • Educational level • Income • Country of birth	-	-	-	

Social Insurance Agency

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	• Age • Gender	-	• Disability and sick leave payments	• Disability pension • Sick leave	
Patient administrative systems (claims)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Regional (Stockholm, Västra Götaland, Skåne)	• Age • Gender	• Diagnoses (ICD-10) • Procedures	• Inpatient care • Outpatient visits • Primary care visits	• Hospitalization or visits	

UNITED KINGDOM

Clinical Practice Research Datalink (CPRD)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National (England)	• Year of birth • Gender	• Diagnoses • Referrals (linked to secondary care data)	• Unit costs of primary consultation services (PSSRU "Unit Costs of Health & Social Care") • Hospital costs (NHS reference costs) • Drug costs available through NHS Electronic Drug Tariff, British National Formulary (BNF), NHS Information Centre's Prescription Cost Analysis (PCA), or NHS Dictionary of Medicines and Devices	• Clinical outcomes • Patient-reported outcomes	Can be linked to other datasets if patients consented to linkage: • Hospital Episode Statistics (HES), including imaging data • Death Registration data (Office for National Statistics, ONS) • National Cancer Registration and Analysis Service (NCRAS) • Mental Health Dataset (MHDS) • Measures of relative deprivation at Lower Layer Super Output Area (LSOA) level for practices and patients

Hospital Episode Statistics (HES)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National (England, all hospital	• Age group • Gender	• Diagnoses	• Healthcare resource groups (HRG) data	• Clinical outcomes	e.g. Ruiz et al. PLOS One. 2015;10(12):e014537

episodes)	• Ethnicity • Geographic information	• Procedures (admissions, A&E, outpatient)		• PROs for four procedures (hip replacement, knee replacement, varicose vein, groin hernia surgery)	2.
Primary Care Clinical Informatics Unit (PCCIU)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (Scotland), appr. 300 practices (1/3 of Scottish population), 2000-2011	• DOB (month and year only) • Gender • Postcode	• Patient encounters • Diagnoses • Tests and results • Measurements taken	• Procedures • Prescriptions		No free text collected. No patient or clinician identifiable data collected.
The Health Improvement Network (THIN) database					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (UK, appr. 12.3 million patients from 587 practices, nationally representative)	• Year of birth • Gender • Household identifier • Practice registration date and status • Socio-economic data (post code level)	• Diagnoses • Symptoms • Tests and results • Measurements taken • Referrals to secondary care • Secondary care details: admissions, medication, diagnosis, investigation • Lifestyle data	• Prescriptions • Consultations • Hospital admissions	• Clinical outcomes • Death	Based on electronic medical records. Held by IMS Health (commercial). Linked to HES.

QResearch					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (England, appr. 13 million patients from 1500 practices)	• Age • Gender • Socio-economic status (postcode level)	• Preventive care • Diagnoses • Referrals (linked to secondary care data) • Tests and results	• Prescriptions • Consultations	• Clinical outcomes • Death (linked to ONS data)	Primary care database linked to secondary care data. Data are de-identified at source.
Research One / TPP					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
UK 44 Million patient records from > 2,600 practices and all prisons (142)	• Age (month and year of birth) • Gender • Sector-level post code • Mid-level super output area • Ethnicity • Occupation • Rurality indices • Deprivation indices	• Diagnoses • Appointments • A&E admissions • Hospital admissions • Referrals • Allergies • Vaccinations • Waiting lists	• Prescriptions • Consultations • Some areas primary only, some primary and secondary	• CTV3 Read codes • ICD10 Classification of Diseases • OPCS4 Classification of Interventions and Procedures • A&E Diagnosis, Treatment and Investigation codes • Death (date and cause)	Data from primary care practices using SystmOne
Consultation in Primary Care Archive (CiPCA)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Regional (North Staffordshire, 14 practices, since 2000)	• Age • Gender • Socio-economic status (small area deprivation score)	• Diagnoses • Symptoms • Referrals • Investigations	• Prescriptions • Consultations	• Clinical outcomes • Patient-reported outcomes	Uses a unique pseudo-anonymised ID to track patients over time

UK Biobank					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (UK, appr. 500,000 volunteer participants)	• Age • Gender • Socio-economic data (self-reported) • Environment, housing (self-reported)	• Health behaviour • Tests • Genetic information • Biosamples • Cognitive function • Hearing • Imaging • Accelerometry	• ?	• Clinical outcomes (linked to primary care, secondary care, and cancer registries) • Death • Patient-reported outcomes	e.g. Emerging Risk Factors Collaboration et al. JAMA. 2015;314(1):52-60.

SAIL databank					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (Wales)	• Age (week of birth) • Gender • Geographic information (Lower Super Output Area)	• Diagnoses (primary care and inpatient) • Test results • Interventions /operations • Referrals • Primary care consultations • A&E attendance • Critical care data • Outpatient hospital appointments • Inpatient visits	• Prescriptions • Primary care consultations • A&E attendance • Critical care details (incl. intensity of care) • Outpatient appointments • Inpatient visits and procedures	• Death • Clinical outcomes	Dataset linking birth and death registers with outpatient and inpatient (including detailed critical care) care and demographic data. Can be further linked (but requires separate permission) to: • Surveys (Active Adults Survey; Welsh Health Survey; National Survey for Wales) • Disease registries (cancer registry; bowel screening; breast screening; cervical screening; congenital anomaly register)

Mental Health Services Data Set (MHSDS)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (England)	• Age • Gender • Post code • Ethnicity • Marital status • Religion • Language • Accommodation • Employment	• Diagnoses • Referrals • Contacts with carers	• Care plans • Hospital admissions	• Clinical outcomes	
CALIBER					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (UK, appr. 10 million patients)	• Age • Gender • Ethnicity • Deprivation • Marital status	• Health behaviour data • Diagnoses • Primary care consultations • Hospital admissions • Test results	• Prescriptions • Admissions • Primary care consultations	• Clinical outcomes (linked CPRD, HES and MINAP) • Death (including cause of death)	Dataset linking primary, secondary, and registry data, with a focus on CVD.
Scottish health records (electronic Data Research and Innovation Service, eDRIS)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Scotland		Primary and secondary care			
Secondary care prescribing (innovation score card from NHS Innovation Health and Wealth Strategy)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
		Statistics on hospital prescribing			

		of recently approved NICE technology appraisals			
Mortality data (ONS)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
England and Wales (tbc)	Age Gender	Annual deaths with cause of death			
Survey data source (Millennium Cohort Study)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
United Kingdom	Age Socio-economic background Gender Other personal information				
Survey data source (General Practice Patient Survey)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
England	Age Ethnicity Gender				
Survey data source (Labour Force Survey)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
United Kingdom					

4.2 Disease-specific data sources: multiple sclerosis

FINLAND					
MS disease register					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
By April 2016 all 5 university hospitals and 6 central hospitals have joined the register					

FRANCE					
EDMUS: European Database for Multiple Sclerosis and other related diseases					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
European, public					
French Multiple Sclerosis Registry (OFSEP)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
French Multiple Sclerosis Registry (OFSEP)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**

HUNGARY

Csongràd County MS registry

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
Since 1996. In addition, there are apparently 19 Hungarian multiple sclerosis centres, each with a disease registry. All this needs to be confirmed					

ITALY

National registry

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
Italy (≈45,000 patients currently registered coming from the main 62 MS centres at national level, but	• Gender Date of birth Place of birth Residence Fiscal code MS familiarity	• Diagnosis time Mc Donald classification Relapse time/type Disability scores (Expanded Disability Status Score – EDSS, Functional System Score	Specialist visits Diagnostic tests (mainly magnetic resonance	• Pharmacovigilance Clinical bio-markers Costs Quality of life	

aiming at involving 130 centres in the next future). 2015-now (a smaller registry was already active since 2001)		– FSS, 9-Hole Peg Test – 9-HPT, Timed 25-Foot Walk - T25-FW, Paced Auditory Serial Addition Test -PASAT3) Treatments (and reason for interrupting)	imaging) Drug consumption Laboratory exams		
Regional registries					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Tuscany (≈2300 patients currently registered). 2006 – now. Individual	n/a	n/a	n/a	n/a	
Liguria (≈680 patients currently registered). 2014 – now. Individual.	n/a	n/a	n/a	n/a	
Sicily. 2017 – now. Individual.	n/a	n/a	n/a	n/a	
Italy (≈45,000 patients currently registered coming from the main 62 MS centres at national level, but aiming at involving 130 centres in the next future). 2015 – now (a smaller	Gender Date of birth Place of birth Residence Fiscal code MS familiarity	Diagnosis time Mc Donald classification Relapse time/type Disability scores (Expanded Disability Status Score – EDSS, Functional System Score – FSS, 9-Hole Peg Test – 9-HPT, Timed 25-Foot Walk - T25-FW, Paced Auditory Serial Addition Test - PASAT3) Treatments (and	Specialist visits Diagnostic tests (mainly magnetic resonance imaging) Drug consumption Laboratory exams	Pharmacovigilance Clinical bio-markers Costs Quality of life	

registry was already active since 2001). Individual.		reason for interrupting)			

NORWAY

National quality register for Multipel Skleros

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National, Time coverage: 2001-2017. MS-biobank from 2006.	· Age · Gender · Family members with MS	· Diagnoses (ICD-10) · CSF-analyses · MRI at diagnosis	· Ongoing treatments · Medications	· Number of relapses · MRI examination · EDSS · Multiple Sclerosis Functional Composite (MSFC) · Symbol Digit Modalities Test (SDMT) · Symptoms To be included in electronic registration system: o Fatigue Severity Scale (FSS) o Multiple Sclerosis o Impact Scale (MSIS-29) o EQ-5D	

SWEDEN

National quality register for Multiple Sclerosis (SMSreg)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	· Age · Gender ·	· Expanded	· Number of visits ·	· Date of death ·	

(80% coverage) Time coverage: 2001-2017	Family history of MS · Occupation	Disability Status Scale (EDSS) · MSSS-score · Date of diagnosis · Basis for diagnosis · Treatment · Relapses · Date of relapse · MRI examinations · Laboratory analyses · BMI · Cerebrospinal fluid · Function scales · EQ5D · Work capacity · Rehabilitation	Care provide	Date of relapse · Adverse events (type and degree) · EQ5D · EDSS	

UNITED KINGDOM

MS Register

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
United Kingdom	Age Gender Ethnicity Family background Other personal information	Clinical study + online survey = MS Register Data from: Online questionnaires completed by volunteer's Clinical study NHS routine data			Data can be anonymously linked

Scottish Multiple Sclerosis Register (Information Services Division of NHS Scotland)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
Scotland	Gender Ethnicity Other demographic information	Demographic Family history of MS Date of first symptoms Referrals MS nurse patient involvement after initial diagnosis of MS Types of investigation Diagnosis			

Scottish health records (electronic Data Research and Innovation Service, eDRIS)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
Scotland		Primary and secondary care			

LSE IMPRESS (International Multiple Sclerosis Study)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Europe	Age Gender Marital status Education Employment status Other personal information	demographic, disease variables, Disease Modifying Drug (DMD) consumption, healthcare and informal care utilisation, productivity losses, QoL, Physical disability, experience with MS (treatment satisfaction, future treatment expectations, caregiving arrangements and sources of information for MS.			
European Register for Multiple Sclerosis (EUreMS)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Europe					
MS Base					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
27 countries worldwide					

Survey - New insights into the burden and costs of multiple sclerosis in Europe					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Europe	Age Gender Living arrangements Education situation Work situation	Disease, health-related quality of life, demographics, inpatient care, outpatient care, equipment, community and family assistance			Cross-sectional, observational study in 16 countries. Uses a questionnaire

4.3 Disease-specific data sources: inflammatory bowel disease

FRANCE

French cohort of Inflammatory Bowel Disease (IBD)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National					

HUNGARY

HUPIR (Hungarian Pediatric IBD Registry) is a national registry for paediatric patients managed by 44 paediatric gastroenterologists in Hungary

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	• Paediatric				

ITALY

Regional IBD Disease Registry (Liguria), Since 2014 (data included retrospectively since 2011)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
Regional	• Name/Surname	First visit:	• Outpatient visits	• Incidence rate	

(Liguria)	• Date of birth • Gender • Citizenship • Residence • Number of children • Educational level • Job status/type	• Main diagnosis (Crohn Disease or Ulcerative Colitis or Indeterminate Colitis), Family profile, Personal physiological profile (e.g. body mass index, smoking status), Personal pathological profile (IBD- and non-IBD-related) At each visit: • IBD localization, Symptoms, Fistulas/Fissures, Abdominal mass/pain, Number of liquid stools/day, Extra-intestinal manifestations, Harvey-Bradshaw Index/Mayo Score At the occurrence: • Laboratory test results, Drug therapies (e.g. type, administration mode, adverse events), Surgery (e.g. date, type, reason), IBD-related neoplasm (type, treatment), Pregnancies	• Laboratory exams • Diagnostic tests • Drug therapies • Surgery	• Mortality rate • Healthcare costs	
Sub-regional (Local Health Authority for the Forlì Province in the Emilia-Romagna Region of Italy) from 1993-2013, extended to all of the Romagna section of the Emilia-Romagna Region from 2010					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
Sub-regional	• Name/Surname • Fiscal code • Date/place of birth • Gender • Citizenship • Residence • Date/place of immigration or emigration • Date/place of death	• ICD9-CM diagnoses codes (555* or 556*) • SNOMED codes (pathological anatomy) • Date of diagnosis • Certainty of diagnosis (4=SDO data; 5=histological or clinical results) • Celiac disease (yes/no) • Dysplasia (mild/severe)		• IBD case-mix • Incidence rate • Prevalence rate • Incidence of cancer in IBD • Preference-based quality of life data (SF-36) for a subset of patients with possibility to generate QALYs	Linked by individual identification codes to the Romagna Cancer Registry

	• General practitioner				

SWEDEN					
National quality register for Inflammatory Bowel Disease, 2005-2017					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (64% coverage)	• Age • Gender • Family history of IBD	• ICD diagnosis • Date of diagnosis • Basis for diagnosis • Number of visits • Treatment • Test results • Symptoms • Montreal Classification of Crohn´s disease • Montreal Classification of Ulcerative colitis • Biomarker (Faecal calprotectin) • Endoscopy grading • Smoking habits •	• DRG type (medical or surgical) • Number of surgeries • Responsible medical staff •	• Date of death • Date of progression • Adverse events (type and degree)	

UNITED KINGDOM

IBD Registry (British Society of Gastroenterology), www.ibdregistry.org.uk, initiated in 2011.
On 1 March 2017 the Royal College of Physicians' (RCP's) Inflammatory Bowel Disease (IBD) Audit Programme successfully transitioned the biologics audit to the UK IBD Registry. https://www.rcplondon.ac.uk/projects/ibd-programme

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National	• Year of birth • Gender • Socio-economic data (through post code level •	• Diagnosis • Medication • Biologics • Hospital admissions • Surgery • Site of disease • Severity scores	•	• IBD Impact • Days lost • PROs •	Current focus on IBD patients on biologics Supported by a range of IBD domain stakeholders

4.4 Disease-specific data sources: lung cancer

FINLAND					
The Finnish Cancer Registry					
Coverage	• **Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (1953-2017). Lung cancer is included	• Age • Gender	-	-	-	
The Mass Screening Registry					
Coverage	• **Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
National (1968-2017). Lung cancer is included	• Age • Gender	-	-	-	

FRANCE

France Cancer Registry

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
Multiple regional cancer registries	• Individual code	-	-	-	

HUNGARY

Cancer Registry

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National (1999-2017). Includes Lung cancer	-	• Date of death (cause of death not available, although it is collected)	-	• Date of death • Mortality rates	

ITALY

Cancer registries

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
38 cancer population registries (at municipal, provincial or regional level). About 34 million	• Individual-level: • Name • Age • Gender • Home address • Fiscal code	• Cancer localization • Cancer stage • Cancer histological type • Cancer biomarkers • Hormone receptor status • Screening status	-	• Cancer incidence/prevalence/mortality • Screening/treatment effectiveness • Environmental, lifestyle, work-related and genetic risk factors • Socio-economic and geographical inequalities in	

citizens (57% of the total Italian population The first registry (city of Varese) was instituted in 1976, the last ones in 2016		• Clinical conditions • Treatments • Other information according to the registry		access to care and incidence/mortality	

NORWAY

National quality register for lung cancer

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National (97 %). 2013-2017	• Age • Gender	• Diagnoses (ICD-10)	• Inpatient care • Outpatient visits • Primary care visits	• Lung cancer • Medications • Radiotherapy	• ECOG • Date of death

SWEDEN

National quality register for lung cancer

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
National (88% coverage). Since 2002	• Age • Gender	• Diagnoses (ICD-10) • Date of diagnosis • Basis for diagnosis • Tumour stage • EGFR and ALK • Treatment	• Diagnostics • Treatments • Waiting times • Follow-up	• Date of death • Date of progression • Adverse events (type and degree)	

UNITED KINGDOM

National Cancer Registration and Analysis Service (PHE)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
England	• Name • Address • Age • Sex • Date of birth • NHS number	• Diagnosis • Treatment • Outcomes	• Treatments	• Date of death • Cause of death Patient reported outcome measures and patient experience	Responsible for all cancer registration in England. 8 regional offices. NCRAS accesses data from a range of sources including HES, pathology, radiology, ONS, COSD, cancer waiting times, and patient administration systems. Data sources are combined to produce, for each patient, pathway completed registration dataset. CancerStats is an online portal by NCRAS.

Cancer Outcomes and Services Dataset (COSD)

Coverage	Demographic	Clinical	Resource use	Outcomes	Comments
England, National coverage since	• Name • Address	• Demographics • Referrals	• Treatments • Procedures	• Death	National standard for reporting cancer

2013	• Age • Sex • Date of birth • NHS number	• Imaging • Pathology • Diagnosis • Care plan • Treatment		• Patient reported outcome measures and patient experience	in NHS in England since January 2013. Replaced Cancer Registry dataset. About 200 data items for lung cancer. Specifies the items service providers submit monthly to NCRAS. Compilation of many different sources into one complete patient pathway record.
Cancer incidence and survival statistics (ONS, ISD Scotland, WCISU, N Ireland Cancer Registry)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
England, Scotland, Wales, Northern Ireland	• Age • Sex	• Survival • Cancer registrations		• Survival • Mortality	
National Cancer Patient Experience Survey (NHS England)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
England. Started 2010. 71,000 patients took part in 2016	• Year of birth • Sex • Sexual identity • Long-standing conditions • English language • Ethnicity	• Seeing GP • Diagnostic tests • Finding out what was wrong with you • Deciding the best treatment for you		• None (patient experience with care provided only)	Monitors progress on cancer care nationally. National, hospital trust and CCG level.

		• Clinical nurse specialist • Support for people with cancer • Operations • Hospital care as an inpatient • Hospital care as an outpatient • Home care and support • Care from your GP • Your overall NHS care • Your condition			
Systemic Anti-Cancer Therapy (SACT) dataset (PHE)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
England	• NHS number • Date of birth • Gender • Ethnicity • Postcode	• Patient and tumour characteristics • Trust and consultant details • Treatment characteristics including drug names and drug combinations	• Treatment	• Regimen outcome summary • Outcome fields	Collects information reported routinely by NHS trusts about therapy activity from NHS England chemotherapy providers Integrated with other clinical datasets from the NHS
Diagnostic Imaging Dataset					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
	• NHS number • Date of birth • Gender	• Referral • Test • GP practice			Information about imaging tests and

	• Ethnicity • Postcode	• Where patient came from (referred, outpatient etc) • Waiting times			scans from radiology departments in NHS hospitals and NHS-funded activity in private providers. Can be linked to other datasets to understand link with diagnoses.
National lung cancer audit (Royal College of Physicians)					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
England, Wales.	• NHS number • Birth date • Postcode • Gender • Ethnicity • Date of diagnosis	• Process • Diagnosis • Treatment • Pathology	• Treatments	• Cancer staging	Data submitted by all trusts.
National Radiotherapy Dataset					
Coverage	**Demographic**	**Clinical**	**Resource use**	**Outcomes**	**Comments**
England.	-	• Radiotherapy attendance • Attendance identification • Radiotherapy episode • Prescription • Exposure	• Prescriptions Radiotherapy procedures	-	Monthly data from NHS providers of radiotherapy services in England

DO→IT

innovative medicines initiative

Big Data for Better Outcomes

www.bd4bo.org

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 116055. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA

efpia