RAND EUROPE

Reuse of health data by the European pharmaceutical industry

Current practice and implications for the future

Lucy Hocking, Sarah Parks, Marlene Altenhofer, Salil Gunashekar

For more information on this publication, visit www.rand.org/t/RR3247

Published by the RAND Corporation, Santa Monica, Calif., and Cambridge, UK

RAND Europe is a not-for-profit research organisation that helps to improve policy and decision making through research and analysis. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

This report was prepared for the European Federation of Pharmaceutical Industries and Associations (EFPIA).

Preface

The reuse or secondary use of health data refers to the use of data initially collected for some other primary purpose (e.g. as a by-product of the care process or initially for other types of research or administrative purposes). Against the backdrop of a rapidly evolving health data landscape, the European Federation of Pharmaceutical Industries and Associations (EFPIA) commissioned RAND Europe to undertake a focused review study to explore current practices related to the reuse of health data by the European pharmaceutical industry. Specifically, the objectives of this study were to understand: (i) how different types of health data are reused by the pharmaceutical industry and reasons for this; (ii) the key enablers and barriers to effective reuse of data; and (iii) considerations and potential implications for future action by different stakeholders (including industry and policymakers).

We adopted a multi-method approach to carry out the research that included a targeted literature review, case vignettes exemplifying industry reuse of health data, in-depth interviews with stakeholders across the health data ecosystem and a synthesis workshop.

As the use of health data proliferates and there is a growing recognition of its potential opportunities and also challenges, this report will be of interest to policymakers, regulators, the pharmaceutical industry, health data innovators and researchers, as well as more broadly to anyone interested in the development of health data ecosystems.

RAND Europe is a not-for-profit policy research organisation that helps to improve policy and decision making through research and analysis.

For further information about RAND Europe and this document please contact:

Dr Salil Gunashekar (Research Leader)
RAND Europe, Westbrook Centre, Milton Road
Cambridge CB4 1YG
United Kingdom
Telephone: +44 (1223) 353 329
Email: sgunashe@rand.org

Executive summary

Introduction and objectives of the study

There is a growing recognition of the potential benefits of reusing health data, while at the same time it is acknowledged that there are numerous technological, regulatory, economic and social challenges that need to be addressed in order for stakeholders to capture value from health data and create enabling health data ecosystems. Against this backdrop, the European Federation of Pharmaceutical Industries and Associations (EFPIA) commissioned RAND Europe to undertake a focused review to explore current (and emerging) practices related to the reuse of health data by the European pharmaceutical industry. Specifically, the objectives of this study were to understand: (i) how different types of health data are reused by the pharmaceutical industry (e.g. what are they being used for) and reasons for this; (ii) the key enablers and barriers to effective reuse of data; and (iii) considerations and potential implications for future action by different stakeholders (including industry and policymakers).

We define reuse or the secondary use of health data as the use of health data initially collected for some other primary purpose, where health data is any 'personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status' (European Union 2016, L 119/34). We broadly considered four types of health data: electronic health records (EHRs), registry data, health systems data and clinical trial data.

Methodology and caveats

In order to meet the objectives of the study, we adopted a mixed-methods approach involving:

- A targeted literature review, focusing on articles between 2009 and 2018 inclusive describing the reuse of health data by the pharmaceutical industry.
- Identification and development of a set of 12 case vignettes illustrating how the European pharmaceutical industry reuses health data in practice (see Box 1 below).
- A series of interviews with stakeholders across the health data ecosystem.
- A synthesis workshop to reflect on the findings and identify potential implications for future action by stakeholders involved in the health data ecosystem.

There are a few caveats that should be borne in mind when interpreting the findings presented in this report. First, there has been a rapid increase in the reuse of health data by the pharmaceutical industry in recent years, however, much of what industry has done so far, and is currently doing in this space, has not been published. To mitigate against this limitation, we used interviews to engage with a number of pharmaceutical industry representatives. Second, the focus of this study was on the reuse of health data by

the European pharmaceutical industry. We are aware that most large pharmaceutical companies operate in several countries, however, we focus on examples within Europe. Finally, this report is not intended to be a comprehensive overview of all secondary analysis activities undertaken by the European pharmaceutical industry. Rather, it should be viewed as providing a current snapshot of the reuse of health data by industry.

Box 1: Case vignettes illustrating reuse of health data by the European pharmaceutical industry

CV1.	Use of Virtual International Stroke Trials Archive (VISTA) data to develop a method to improve the precision of cost-effectiveness studies related to stroke trials.
CV2.	Use of the Clinical Practice Research Datalink (CPRD) in the United Kingdom to investigate channelling bias in the populations prescribed glucose-lowering drugs.
CV3.	Use of the THIN database, an EHR, to identify adverse side effects of drugs.
CV4.	Use of non-randomised study data and observational data in network meta-analyses of randomised control trials (RCTs) to assess schizophrenia and in-stent restenosis treatment effects.
CV5.	Exploration into the feasibility of using Swedish EHRs and health registries to supplement or replace control populations in RCTs for asthma.
CV6.	Using paediatric pulmonary arterial hypertension health registries to improve understanding of the disease.
CV7.	Using the Fabry Outcome Survey registry to explore the impacts of the disease on female reproductive and pregnancy outcomes.
CV8.	Use of the THIN and Humedica EHR databases to explore any associations between BMI and an increased risk of non-alcoholic fatty liver disease.
CV9.	Using the Schizophrenia Outpatient Health Outcome dataset to explore the methods of identifying drug effect modifiers for schizophrenia.
CV10.	Using EHRs to identify negative symptoms of schizophrenia and explore if these are associated with clinical outcomes.
CV11.	Testing a data derivation tool to identify individuals with type II diabetes across four different types of health data sources.
CV12.	Using EHRs to obtain NICE reimbursement approval for a lung cancer treatment.

How is the pharmaceutical industry reusing health data?

What types of health data are being reused by the pharmaceutical industry?

EHRs, health registry data and clinical trial data are the most frequently mentioned types of health data being used by the European pharmaceutical industry. EHR data are used particularly frequently as there are established databases to access these (e.g. the UK's CPRD). Other types of health data used by the European pharmaceutical industry include biobank data, prescribing and dispensing data, and claims data. More recently, less traditional types of real-world data are also being used, e.g. social media data or data from wearables such as health monitoring devices and smartwatches.

Why is the pharmaceutical industry reusing health data?

There are a variety of reasons the pharmaceutical industry is reusing health data including:

- Real-world data enable better insights into the real world than an artificial setting (such as clinical trials) could.

- Secondary health data analyses lead to more efficiency and help reduce costs for both industry and health systems.
- The growing availability of health data is contributing to their increasing use.
- Secondary analyses can allow for control groups to be constructed from data that already exists, rather than needing to enrol patients.

What is the pharmaceutical industry reusing health data for?

The pharmaceutical industry is reusing health data across the research and development pathway.

At the discovery and drug development stage, real-world data are used:
• To identify diseases or indications of a significant burden to a wider population. • To better understand a disease, e.g. the impacts of a disease on the wider health and well-being of patients, risk factors associated with a disease or disease progression. • To understand the prevalence of a disease or condition. • To provide new insights into disease associations or comorbidities and therefore to target new populations and indications for future research. • To develop targeted and personalised therapies and drugs. • To develop new analytical methods.
At the clinical research stage, real-world data are used:
• To inform clinical trial design, e.g. to improve the study population selection for clinical trials, to predict the number of potential patients or to assess the efficacy of a new drug. • To create new approaches to patient stratification. • In feasibility studies. • Alongside or instead of control groups for trials to reduce the need to enrol patients as controls.
At the marketing authorisation and market access stage, health data are used:
• For medicine authorisation and regulatory purposes. • To support market access discussions, e.g. to conduct health technology assessments (HTA), identify how competitive drugs are used on the market and to support pricing discussions. • To conduct cost-effectiveness analyses.
At the post-authorisation stage, health data are used:
• To support pharmacovigilance, i.e. to identify safety issues and adverse reactions. • For pharmacoepidemiology, i.e. to understand treatment effects across patient populations, to identify patient groups resistant to drugs, as well as to get insights into patient adherence. • To add to the medical evidence base and inform changes in practice guidelines. • To support effectiveness comparisons between new drugs and existing drugs. • To inform drug repurposing, i.e. the identification of diseases and conditions that could be treated with an existing drug.

To what extent is the pharmaceutical industry reusing health data?

The use of health data by the pharmaceutical industry has become more commonplace over the past few years; there has been a particular increase in its use at the discovery and drug development stages whereas previously it was mainly used at the market entrance and post-authorisation stages.

The scale of real-world data use differs by the type of pharmaceutical company: European companies developing new drugs are using real-world data more extensively than, for example, manufacturers of

generic drugs. In general, larger pharmaceutical companies tend to use real-world data more often than smaller ones.

How does the pharmaceutical industry access health data and how are these data safeguarded?

Data are accessed by the pharmaceutical industry in a variety of ways:

- Collaborations between industry and other organisations (e.g. other pharmaceutical companies, healthcare providers, universities and research organisations, private organisations, regulators, and policymaking and arm's-length bodies) help provide industry with access to health data.
- Industry uses publicly-available datasets as well as buys health data from vendors.
- The ease of accessing health data depends on the type of health data and the nature of the study they are intended for, as well as other contextual factors such as country.
- There is a spectrum of governance arrangements for access to health data, from publicly available (anonymous) datasets to completely restricted access.
- Access to health data is a particular challenge when a study requires multiple datasets, as this requires research teams to approach each dataset owner individually regarding access.

There are several ways how the use of health data by industry is safeguarded: safeguards set up by the data owner; internal governance and control policies; restricted access to data within an organisation; ethical approvals; confidentiality agreements; privacy legislation; using the same security standards as used for clinical trials; and policies restricting access such that analyses can only be carried out by the data owner or by specific organisations that have oversight of the data.

What are the potential impacts of the secondary use of health data?

The literature and interviewees reported more often on the potential positive impacts of the use of health data than on negative impacts. Several interviewees felt that there were no negative impacts at all.

Identified potential positive impacts include: improved treatments for patients; accelerated research and drug development, and therefore also accelerated treatment access; better understanding of and improvement of treatment efficacy; improved pharmacoepidemiology and pharmacovigilance; the most appropriate patient populations can be recruited; if research and development processes are accelerated, industry can get their products earlier on the market; and secondary analyses can support pricing discussions, and thus help companies get reimbursement contracts for their drugs.

Interviewees did not refer to concrete examples of negative impacts of reuse of health data that they had observed, but only to potential ones (note that this does not indicate that there are no negative impacts associated with the reuse of health data). In terms of potential negative impacts of reuse of health data, it was mentioned that this could occur if pharmaceutical companies broke personal data laws or if they used the data for purposes other than the original purposes. This could lead to less trust and restricted access to health data. Negative impacts of reuse of health data could also occur if pharmaceutical companies conducted poor-quality analyses, e.g. if they 'fished' for the results they want to find; and if health data have low quality, this could lead to poor research results and variable evidence.

What are the enablers and barriers to reusing health data?

We identified five themes which describe both the enablers and the barriers of reuse of health data by the pharmaceutical industry. Below we describe the barriers and enablers, grouped under these five themes.

Access, quality, accuracy and standardisation of health data

Reuse of health data by the pharmaceutical industry is hindered by:

- Restrictions on data, meaning that it is not accessible or only accessible to academic research teams.
- Poor quality data, i.e. data containing missing data and inaccuracies.
- Available data not containing all the variables required for analysis or not being updated as regularly as might be desired (this can occur because the primary purpose of collecting data was not for analysis).
- Unintentionally biased data due to the primary purpose of collecting the data not being for its analysis.
- Lack of standardisation and interoperability across datasets meaning analyses using multiple datasets is difficult.

Reuse of health data by the pharmaceutical industry is enabled by:

- Data that are easily accessible to the pharmaceutical industry, e.g. by means of governance procedures that support access.
- Data that are relatively inexpensive to access (this can influence which data are used, e.g. EHR data can be cheaper than prescription and claims data).
- Partnerships with data holders or other bodies who are able to access data.
- The availability of high-quality, curated datasets.
- Secondary analysis tending to cost less than traditional analyses, e.g. secondary analysis can be cheaper than conducting a randomised clinical trial.

Administrative factors and collaboration

Reuse of health data by the pharmaceutical industry is hindered by:

- Administrative factors associated with project structure and management, e.g. staff turnover and slow decision making, particularly on collaborative projects across multiple locations.
- Upfront start-up costs to acquire the skills and infrastructure needed to effectively reuse health data (these costs can disincentivise senior management from investing in reuse of health data).

Reuse of health data by the pharmaceutical industry is enabled by:

- Effective collaborations with a range of stakeholders, both to provide access to data and to provide different ideas, perspectives and skills.
- Cultures within pharmaceutical firms that are open to and accepting of the value of the reuse of health data

Regulations and guidelines (including issues related to data protection and data privacy)

Reuse of health data by the pharmaceutical industry is hindered by:

- A lack of clear and uniform regulations on and information about what is and is not acceptable in terms of health data reuse. (This includes whether secondary analysis can be submitted as valid evidence for a drug's efficacy, how health data are allowed to be used by the pharmaceutical industry, and what is legal and acceptable in terms of secondary analysis.)
- The existence of intra- and inter-country differences in regulations and guidelines for secondary analysis.
- HTA bodies not recognising evidence generated through secondary analysis of health data.
- Lack of clarity on ownership of outputs of secondary data analysis, and who has the right to share data and with whom.

- Lack of clarity in relation to GDPR and the systems that need to be in place to ensure effective data governance and protection of health data.
- Only being able to access anonymised health data which may have had desired variables removed.

Reuse of health data by the pharmaceutical industry is enabled by:

- Regulators and policymakers recognising the value of and demanding real-world evidence before they make decisions.

Data analysis skills and capabilities

Reuse of health data by the pharmaceutical industry is hindered by:

- Lack of development of methods for reusing health data.
- Lack of skills and experience at pharmaceutical companies to effectively analyse health data.

Reuse of health data by the pharmaceutical industry is enabled by:

- Continual development of new tools and methods to reuse health data.
- Pharmaceutical companies having access to staff – either in-house or externally – with adequate skills to conduct high-quality analyses, interpret the data correctly and/or judge the accuracy of health data.

Public and healthcare provider trust

Reuse of health data by the pharmaceutical industry is hindered by:

- Lack of public and healthcare provider trust due to concerns over private companies accessing personal health data and not having adequate data protection processes in place.
- Concerns in pharmaceutical companies over lack of public trust and potential loss of public trust.

Reuse of health data by the pharmaceutical industry is enabled by:

- Clear governance processes detailing how data can be used and who it can be used by.
- Only using anonymised datasets in the secondary analyses of health data.

Reflections on the future of the reuse of health data

What might the future look like?

- It is likely that secondary analysis of health data will increase in the near- to medium-term future (one to five years in the context of this study), expanding to cover different disease areas, new analytical methods and different types of data.
- As they become more involved in the health data ecosystem, patient organisations are likely to play a bigger role in secondary data analysis in the future, which may help overcome some data protection concerns held by the public.
- Regulations and guidelines associated with health data reuse are likely to become clearer and more coordinated in the future, as has been seen recently in reports produced by the US FDA and the EMA.

What are the priority topics for further discussion that might help create a sustainable ecosystem in which health data is reused effectively?

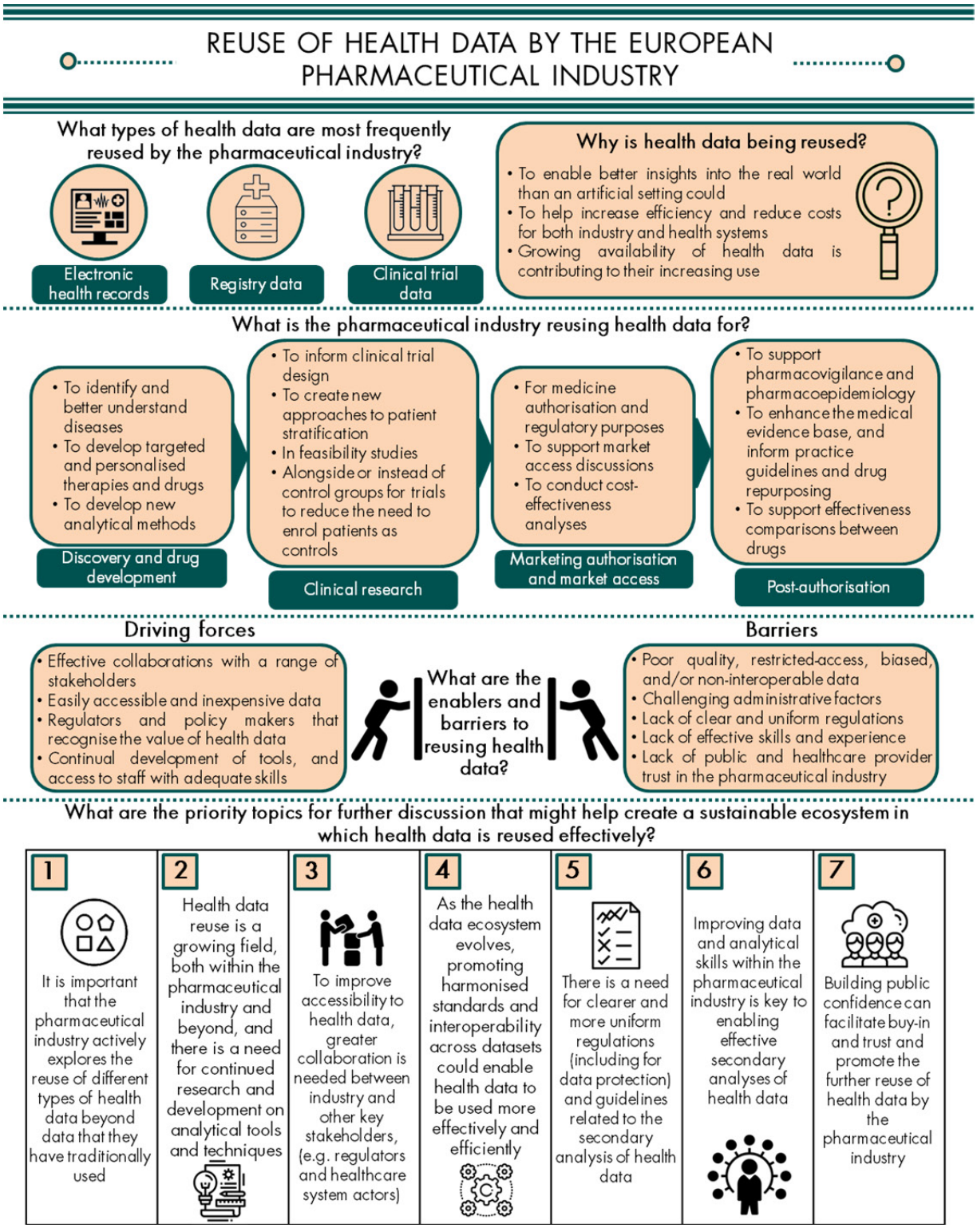
Based on the findings described above, we proposed a set of wide-ranging ideas or topics which need to be considered by the pharmaceutical industry and other stakeholders to get the most out of the reuse of

health data, and in doing so, help create an effective and enabling health data ecosystem. These are the following:

- It is important that the pharmaceutical industry actively continues to explore the reuse of different types of health data beyond the data types (such as EHRs, registry data and clinical trial data) that have been traditionally used in secondary analysis.
- Health data reuse is a growing field, both within the pharmaceutical industry and beyond, and there is a need for continued research and development on analytical tools and techniques (including the ability to link different datasets).
- To improve accessibility to health data, a greater degree of collaboration is needed between the pharmaceutical industry and other key stakeholders, such as regulators and healthcare system actors.
- As the health data ecosystem evolves, promoting harmonised standards and interoperability across datasets could enable health data to be used more effectively and efficiently.
- There is a need for clearer and more uniform regulations (including for data protection) and guidelines related to secondary data analysis.
- Improving data and analytical skills within the pharmaceutical industry is key to enabling effective secondary analyses of health data.
- Building public confidence can facilitate buy-in and trust and promote the further reuse of health data by the pharmaceutical industry.

In Figure 1 we provide a visual summary of the project.

Figure 1: Project summary infographic



Source: RAND Europe analysis

Table of contents

List of figures

List of tables

List of boxes

Abbreviations

ABPI	Association of the British Pharmaceutical Industry
AI	Artificial intelligence
BMI	Body mass index
CPRD	Clinical Practice Research Database
EFPIA	European Federation of Pharmaceutical Industries and Associations
EMA	European Medicines Agency
EHR	Electronic Health Record
EMIF	European Medical Information Framework
ENCR	European Network of Cancer Registries
EU	European Union
EU-CTR	EU Clinical Trials Register
FDA	Food and Drug Administration
FIND	Findable, accessible, interoperable and reusable
FOS	Fabry Outcome Survey
GDPR	General Data Protection Regulation
GP	General practitioner
HTA	Health technology assessment
HU	Health Utility
ICTRP	International Clinical Trials Registry Platform
IMI	Innovative Medicines Initiative
INAGEMP	National Institute on Population Medical Genetics
ISAC	Independent Scientific Advisory Committee
MHRA	Medicines and Healthcare products Regulatory Agency
NAFLD	Non-alcoholic fatty liver disease
NICE	National Institute for Health and Care Excellence

OECD	Organisation for Economic Co-operation and Development
PAH	Pulmonary arterial hypertension
QALY	Quality Adjusted Life Year
R&I	Research and innovation
RCT	Randomised controlled trial
THIN	The Health Improvement Network
UK	United Kingdom
US	United States
VISTA	Virtual International Stroke Trials Archive
WHO	World Health Organization

Acknowledgements

We are very grateful to Brendan Barnes at EFPIA for his guidance and support throughout the study. We would like to thank Gracy Crane and Shahid Hanif for engaging with the study as members of the EFPIA steering group for the study. We are also grateful for the expertise and information provided by the many stakeholders we interviewed over the course of the project; their contributions have been vital for the successful completion of the study. Finally, we would like to thank our reviewers, Advait Deshpande and Brandi Leach, for their constructive comments on this report during the quality assurance process.

1. Introduction

1.1. Objectives of the study

The health data landscape has been evolving over the last few years with a growing recognition of the potential benefits of accessing, sharing and using a variety of health data. At the same time, there are numerous challenges – spanning, for example, technological, regulatory, economic and social issues – to be addressed by different stakeholders to capture value from health data and to create enabling health data ecosystems. Against the backdrop of this changing and complex landscape, the European Federation of Pharmaceutical Industries and Associations (EFPIA) commissioned RAND Europe to undertake a focused review study to explore current (and emerging) practices related to the reuse of health data by the European pharmaceutical industry (henceforth, we use the phrase 'industry' to refer to the pharmaceutical industry). Specifically, the objectives of this study were to understand: (i) how different types of health data are reused by the pharmaceutical industry (e.g. what are they being used for) and reasons for this; (ii) the key enablers and barriers to effective reuse of data; and (iii) considerations and potential implications for future action by different stakeholders (including industry and policymakers).

1.2. What do we mean by reuse of health data in the context of this study?

Under Article 4(15) of the European Commission's General Data Protection Regulation (GDPR), health data or 'data concerning health' refers to any 'personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status' (European Union 2016). Health data that were collected in non-research environments and for purposes other than research are also often referred to as real-world data, which has been defined as 'any data not collected in conventional randomised controlled trials (RCTs). This includes data from existing secondary sources (e.g. databases of national health services) and the collection of new data, both retrospectively and prospectively' (Miani et al. 2014).

The reuse of health data (by industry) refers to the use of data initially collected for some other primary purpose (e.g. as a by-product of the care process or initially for other types of research or administrative purposes). This reuse of health data is also often referred to as secondary use of health data. More specifically, secondary use of health data is defined as using 'personal health information […] for uses outside of direct health care delivery' (Safran et al. 2007), for example, for 'analysis, research, quality and safety measurements, public health, payment provider certification or accreditation, marketing and other business applications' (Safran et al. 2007). Health data are collected from a range of sources, such as

clinical trials, epidemiological studies, healthy volunteer studies, routine health care provisions, post-marketing studies, off-label prescription, reporting of adverse events and prescriptions (ABPI 2007). In this study, we are particularly interested in the secondary use of health data by the pharmaceutical industry in Europe. The study does not cover the reuse of health data by the public sector and/or academia.

1.2.1. What are the different types of health data?

The literature refers to a range of different types of health data being reused for reasons other than their original intention. We broadly consider four main types:

(i) Electronic health records (EHR)
(ii) Registry data
(iii) Health systems data
(iv) Clinical trial data.

The definition of each of these and some illustrative examples are provided in Table 1. Some additional types are described in Chapter 3.

Table 1: Types of health data available for secondary use

Type of data	Definition	Illustrative examples
Electronic health records	'Electronic platforms that contain individual electronic health records for patients and are maintained by healthcare organisations and institutions' (US FDA 2018).	• United Kingdom (UK) Clinical Practice Research Database (CPRD) (Health Economics Research Centre n.d.) • My Health Summary (Sundhed 2016) • Hungary's national health app (Collins 2016)
Registry data	Patient or disease 'registries are organised systems that use observational methods to collect uniform data on populations defined by a particular disease, condition, or exposure, and that are followed over time' (EMA 2018). Health registries often focus on specific diseases or treatments and collect data on the health status of patients and their healthcare over time (NIH 2019).	• European Union Clinical Trials Register (EU-CTR) (EU Clinical Trials Register n.d.) • European Network of Cancer Registries (ENCR) (European Network of Cancer Registries n.d.) • International Niemann-Pick disease registry (INPDR n.d.)
Health systems data	Health systems data can include data on 'spending, hospitals, physicians, pharmaceuticals, prevention, mortality, quality and safety, and prices' (Tikkanen 2017) but also audit,[1] insurance and prescription data.	• Chronic obstructive pulmonary disease audit (Roberts 2014) • Organisation for Economic Co-operation and Development (OECD) collection of health

[1] Audit data can be difficult to define, but in this context we find the definition from the patient registry group PARENT useful: 'An audit is an examination or review that establishes the extent to which a condition, process or performance conforms [to] the predetermined standards or criteria. In a registry, audits may be carried out on the quality of data or completeness of records. The audit can be internal or external. Internal audit is carried out by the registry staff, using a concrete plan and specific indicators to

Type of data	Definition	Illustrative examples
		systems data (Tikkanen 2017) • Achmea Health Database (Pijpers n.d.)
Clinical trial data	Data collected through a 'research study in which one or more human subjects are prospectively assigned to one or more interventions (which may include placebo or other control) to evaluate the effects of those interventions on health-related biomedical or behavioural outcomes' (NIH 2017).	• European Union Clinical Trials Database (EU Clinical Trials Register n.d.) • World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP) (The World Health Organisation n.d.) • Cochrane Library (Cochrane Library n.d.)

Throughout the report, where possible, we specify the type of health data being discussed; interviewees often referred to real-world data and therefore in those instances we use that term.

1.3. Summary of methodology

In order to meet the objectives of the study, we adopted a mixed-methods approach involving: (i) a targeted literature review; (ii) identification and development of a set of case vignettes illustrating how the European pharmaceutical industry reuses health data in practice; (iii) a series of interviews with stakeholders across the health data ecosystem; and (iv) a synthesis workshop to reflect on the findings and identify potential implications for future action by stakeholders involved in the health data ecosystem. Each of these tasks is summarised below; a more detailed description about the methodological approach is provided in Annex A.

1.3.1. Targeted literature review

We conducted a targeted review of the literature to gain an understanding of the types of health data the European pharmaceutical industry are using, how the industry uses these data, the enablers and challenges of reusing health data and the possible implications for pharmaceutical companies and beyond. Based on a search focusing on articles published between 2009 and 2018 inclusive (further details are provided in Annex A), a total of 23 academic papers and grey literature documents were included. The number of articles identified was both a function of the bounded scope of the study and the availability of information in the public domain. Relevant information from these documents was extracted which included: the country(-ies) of focus; the purpose of the document/aim of the research; what type of health data was used; why the data was used; how the data was used; enablers; challenges; and implications for future secondary use of data. Information extracted from these 23 articles was used to inform snowball

assess the most significant sources of error as regards the purpose of the registry. External audit is performed by external personnel, in accordance with pre-established criteria.' (PARENT 2015)

searches and to identify additional relevant literature and to identify potential case vignettes (see Section 1.3.2). In total, 41 documents were analysed.

1.3.2. Case vignettes

We used targeted literature searches and snowballing to identify 12 case vignettes to illustrate how the European pharmaceutical industry currently reuses health data. Literature was eligible for inclusion if a representative of a European pharmaceutical company was listed as an author of the publication. For each case vignette, we extracted the following information: aim of the study; how data was used; enablers of reusing the data; challenges to reusing the data; implications for future research, industry or policy; current/prospective use (i.e. whether the article discusses the current or future use of health data); and any other relevant notes. For each study, we invited the pharmaceutical industry authors to an interview to complement information presented in the publications, obtain deeper insights into study, as well as to clarify and fill any gaps related to the reuse of the data which were not discussed in the literature. In total we interviewed seven individuals across 5 of the 12 case vignettes. In this report, we use unique identifiers (CV1, CV2, CV3, etc.) to reference each of the case vignettes.

1.3.3. Stakeholder interviews

In addition to the case vignettes, we conducted key informant interviews with representatives of European pharmaceutical companies and regulatory authorities to better understand the current secondary use of health data. Potential interviewees were identified in the reviewed literature, from our own professional networks, through targeted Google searches for representatives of the European pharmaceutical industry, and through suggestions made by EFPIA and interviewees themselves. A total of ten interviews (eight pharmaceutical industry and two regulatory authority representatives[2]) were conducted; two case vignette interviewees were also asked the same set of questions in addition to the case vignette-specific ones. The evidence that we synthesised from the interviews has been referenced throughout the report using anonymised unique interview identifiers (INT1, INT2, etc.).

1.3.4. Synthesis workshop

In the final phase of the research, we organised a workshop with EFPIA and some members of the EFPIA steering group for the study to reflect on and test the findings of the analyses as well as to discuss any implications for future action by stakeholders within the wider health data ecosystem.

[2] INT4, INT9

1.4. Caveats of the analysis

There are a few caveats that should be borne in mind when interpreting the findings presented in this report (further caveats are provided in Annex A):

- As highlighted by several interviewees, there has been a rapid increase in the reuse of health data by the pharmaceutical industry in recent years, however, much of what industry has done so far, and is currently doing in this space, is not available in the public domain. As a result, there may be additional relevant and interesting cases of the reuse of health data by the European pharmaceutical industry, but these co[3]uld not be captured by our literature searches. To mitigate against this limitation, we used interviews to engage with a number of pharmaceutical industry representatives. Furthermore, the case vignettes serve as illustrative examples of what is happening 'on the ground' in relation to health data reuse by the pharmaceutical industry. They are not intended to provide a definitive 'mapping' of all the different ways in which the European pharmaceutical industry is reusing health data.
- The focus of this study was on the reuse of health data by the European pharmaceutical industry and we are aware that most large pharmaceutical companies operate in several countries. Due to the scope of the study, we only included articles where a specific focus on Europe (or a European country) was explicitly mentioned, and we only selected articles to serve as the basis of the case vignettes where industry authors were explicitly stated to be located in a European country. Nevertheless, we acknowledge that we may not have covered other relevant examples of the European pharmaceutical industry reusing health data.
- This report is not intended to be a comprehensive overview of all secondary analysis activities undertaken by the European pharmaceutical industry. Rather, it should be viewed as providing a current snapshot of the main developments in the reuse of health data by industry, key enablers and challenges observed, and aimed to identify what needs to be done to improve the reuse of health data by European industry.

1.5. Outline of the report

The rest of this report presents the results of this study. In Chapter 2, we present the 12 case vignettes illustrating specific ways in which the pharmaceutical industry reuses health data. Drawing on the evidence from the literature, case vignettes and stakeholder interviews, Chapter 3 outlines how the pharmaceutical industry is currently reusing health data, including the types of data that are being used, why they are being used and what they are being used for. Looking across the different sources of evidence, Chapter 4 discusses the key enablers and barriers for current reuse of health data by the pharmaceutical industry. Finally, in Chapter 5, we provide some concluding remarks based on the

[3] In the case vignettes, in order to retain anonymity, we do not use interviewee identifiers. In the introduction to the case vignettes in which use interview data, we have noted that an interview has contributed to the vignette.

findings from the analysis and articulate potential implications for future action by stakeholders involved in the health data ecosystem.

2. Case vignettes illustrating how the pharmaceutical industry in Europe is reusing health data

This chapter contains 12 case vignettes highlighting a range of ways in which the pharmaceutical industry has reused health data in practice. The majority of the case vignettes (11 out of 12) were identified through a literature search (described in Section 1.3 and Annex A); an additional case vignette was developed based on an interview. The case vignettes are largely based on the publication(s) extracted from the literature describing the study, with some vignettes also including insights from interviews where we were able to arrange interviews with authors of publications. In this chapter, within each case vignette we refer to the work being described as a 'study' and those who carried it out as the 'study team'.

Case vignettes identified from the literature were selected on the basis that at least one author is from the pharmaceutical industry. In practice, this can mean a variety of ways of working including the following or combinations of the following: the pharmaceutical company itself is doing the analyses (possibly with other organisations); the pharmaceutical company is funding the study and the analyses are being conducted by collaborators; the pharmaceutical company is providing steering for the project and the analyses are being conducted by collaborators; and the pharmaceutical company is providing the data and support for a study conducted by others. In the case vignettes, in order to retain anonymity, we do not use interviewee identifiers. In the introduction to the case vignettes in which use interview data, we have noted that an interview has contributed to the vignette. Annex A contains the extraction template used to extract relevant data from the publications that were reviewed for the vignettes, and the template itself that was used to write up the case vignettes.

2.1. Case vignette 1: Use of Virtual International Stroke Trials Archive (VISTA) data to develop a method to improve the precision of cost-effectiveness studies related to stroke trials

This case vignette illustrates the use of archived clinical trial data to estimate Health Utility (HU) values for patients who have had strokes. HU values are used within cost-effectiveness studies within the calculation of Quality Adjusted Life Years (QALYs). The results from this study contributed to improving the ability to carry out accurate cost-effectiveness studies in two ways. First, it illustrates a way to estimate these HU values from the data. Second, it provides values for use within future cost-effectiveness studies for acute stroke trials.

This case vignette is based on Ali et al. 2017.

2.1.1. Who was involved in the study?

The study was a collaboration between staff at the biotechnology company Genentech, five universities (the University of Glasgow, Queen Mary University of London, Glasgow Caledonian University, the University of Nottingham and the University of Washington) and two hospitals (Queen Elizabeth University Hospital and Glasgow Royal Infirmary) (Ali et al. 2017).

2.1.2. What were the aims?

The aim of this study was to estimate HU values for patients who have had strokes. HU values are used within cost-effectiveness studies within the calculation of QALYs.

2.1.3. What data were used?

This study used data from the Virtual International Stroke Trials Archive (VISTA) (VISTA n.d.), an archive of clinical trial data. The aim of VISTA is to provide a central database for stakeholders, including the pharmaceutical industry, to access archived, anonymised clinical trial data related to strokes (Virtual Trials Archives n.d.).[4] The pharmaceutical companies GlaxoSmithKline (GSK), AstraZeneca, Boehringer-Ingelheim and Janssen and Bayer have contributed clinical trial data to the datasets (Virtual Trials Archives n.d.).[5]

2.1.4. Why and how did they use the data?

The study team used data from the VISTA database to enable the development of HU stratified by level of disability. This level of stratification of the UK could not have been calculated without EHR data (Ali et al. 2017).

[4] Specifically, clinical data can be found on acute ischaemic stroke, stroke rehabilitation, intracerebral haemorrhage, stroke prevention, stroke imaging, endovascular, observational stroke data and cognition.

[5] As well as this study, the VISTA database has been used for a range of purposes in stroke research, such as informing future clinical trial design, conducting cost-effective analyses and detecting adverse side effects (Ali et al. 2017; Hesse et al. 2016; Virtual Trials Archives n.d.).

2.1.5. What were the enablers and challenges of the study?

The VISTA database is a collaboration between pharmaceutical companies and the University of Glasgow. Staff that are part of this collaboration at the University of Glasgow thus have the ability to access the high-quality, publicly available clinical trial data without the need for ethical approval (Hesse et al. 2016). Additionally, members of the VISTA steering group are able to contribute data to the archive and thus have influence over which projects get approval to use VISTA (Virtual Trials Archives n.d.). The publication used to produce this case vignette did not list any challenges associated with the use of the data.

2.1.6. What were the safeguards employed to govern the use of the data?

Organisations must apply to gain access to VISTA data which is also anonymised (Virtual Trials Archives n.d.).

2.1.7. What were the potential or realised benefits to patients and public health?

This study aimed to improve cost-effectiveness analyses related to strokes by incorporating HUs which better represent patients.

2.2. Case vignette 2: Use of the Clinical Practice Research Datalink (CPRD) in the United Kingdom to investigate channelling bias in the populations prescribed glucose-lowering drugs

This case vignette illustrates the use of EHRs to investigate how glucose-lowering drugs have been prescribed and whether there is bias in who they are prescribed to. This study is post-drug release research. The results can be used to assess bias in estimates of effectiveness made from real-world data for these drugs in this setting.

This case vignette is based on Ankarfeldt et al. 2017 and an interview.

2.2.1. Who was involved in the study?

This study was a collaboration between staff at the pharmaceutical company Novo Nordisk and two universities (the University of Utrecht and the University of Oxford) and a hospital (the University Medical Center Utrecht) (Ankarfeldt et al. 2017).

2.2.2. What data were used?

The study team used the Clinical Practice Research Datalink (CPRD) (CPRD n.d.), an EHR containing anonymised patient data from a group of general practitioner (GP) practices in the UK linked to other health data. This dataset includes information such as 'diagnoses, mortality, laboratory results, and prescription data' (Ankarfeldt et al. 2017).

2.2.3. What were the aims?

The aim of the study was to understand whether the characteristics of patients being prescribed diabetes treatments differ for different treatments; specifically, whether those being prescribed new diabetes treatments (GLP-1 and DPP-4i) are different from those being prescribed two existing drug classes (basal insulin and sulfonylurea). This is important because if drugs for the same disease are being prescribed to different types of patient, then this can affect the perceived efficacy of the new treatment in the real world. For example, if patients who are prescribed the new treatments are resistant to existing treatments, then these patients may also be more likely to be resistant to the new ones. This phenomenon is called channelling bias.[6]

2.2.4. Why and how did they use the data?

Channelling bias cannot be investigated without real-world data (Ankarfeldt et al. 2017). The CPRD was chosen as the specific data source because the researchers and Novo Nordisk had previous experience using it and knew it contained the type of data they needed. The CPRD is seen as particularly useful for

[6] Channelling bias occurs when drugs with similar treatment effects, often an established and new drug, are prescribed to patients with different prognostics.

diabetes research as it is a database of GP data and, within the UK, GPs are the primary healthcare professionals patients with diabetes come into contact with.

In this study, they extracted a specific population of patients with type II diabetes who were prescribed a glucose-lowering drug from the CPRD database. They then carried out descriptive analysis to explore changes in patient characteristics over time, and outcome analysis to estimate the effect on blood glucose levels and body weight (Ankarfeldt et al. 2017). They found that for these diabetes drugs there was no channelling bias.

2.2.5. What were the enablers and challenges of the study?

This study was made possible by access to the CPRD being granted by the CPRD Independent Scientific Advisory Committee (ISAC) (CPRD 2019) (Ankarfeldt et al. 2017). There was also motivation and interest within the team at Novo Nordisk and Utrecht University to drive the project forward.

One of the challenges faced by the team was that some of the processes within the project, such as accessing the data, took a long time, and hence slowed the project down. This was thought to have been partly due to the lack of commercial benefit for Novo Nordisk for this project, in addition to staff turnover and other factors.

There were also challenges due to data quality; specifically that data for each patient, such as patient's ethnicity and diet, is often not complete. In this study, in order to carry out the analysis, they assumed that missing data was missing at random, and inferred it under this assumption. This assumption may not be correct and could affect the results of the study (Ankarfeldt et al. 2017).

2.2.6. What were the safeguards employed to govern the use of the data?

To access CPRD data the protocol for the work must be approved by the ISAC (CPRD 2019). CPRD data is all anonymised (Ankarfeldt et al. 2017).

2.2.7. What were the potential or realised benefits to patients and public health?

The use of datasets such as the CPRD allow for studies that could not have been carried out before, and may provide more reliable evidence-base on the benefits and risks of drugs. This information may help doctors to prescribe the most appropriate drugs for their patients (Ankarfeldt et al. 2017).

More broadly, this project demonstrated that it is possible to perform effective outcome analysis after a drug has been approved using a comparable set of patients from clinical trial and observational study data.

2.3. Case vignette 3: Use of the THIN database, an EHR, to identify adverse side effects of drugs

This case vignette illustrates the use of The Health Improvement Network (THIN), a UK EHR database, to explore whether EHRs can be used for drug safety assessments. This study specifically aimed to explore whether EHRs can be used to identify the link between drugs and adverse side effects.

This case vignette is based on Cederholm et al. 2015 and two interviews.

2.3.1. Who was involved in the study?

This study was a collaboration between staff at the pharmaceutical companies Eli Lilly, Pfizer, Takeda and Bayer, and the technology and services company Cegedim (Cederholm et al. 2015). As this was an Innovative Medicines Initiative (IMI)[7] research project, the pharmaceutical companies proposed the initial idea and submitted the proposal. Regulators were also involved, including the European Medicines Agency (EMA), the UK Medicines and Healthcare products Regulatory Agency (MHRA) and Swedish regulators, as well as the WHO.

2.3.2. What data were used?

This study used The Health Improvement Network (THIN), an anonymised EHR of GP records for 11 million patients in the UK (although this study only used data from 7.7 million patients) (Cederholm et al. 2015).

2.3.3. What were the aims?

The aim of this study was to explore whether EHRs are a valid tool for detecting adverse side effects associated with drugs earlier than traditional spontaneous reporting of side effects (Cederholm et al. 2015).

2.3.4. Why and how did they use the data?

EHR data analysis provides a potential approach to identifying adverse side effects, as current approaches in identifying these are based on spontaneous reports of adverse effects, which are useful but may not detect more common adverse side effects that have more than one cause (Cederholm et al. 2015). An interviewee noted that EHR datasets are the focus of these analyses as they are seen as more accurate than other data sources, such as claims data, as the information is collected for medical purposes rather than reimbursement purposes, which is the purpose of claims data. The THIN database was chosen in particular primarily because it included the relevant data needed for the research and provided longitudinal data which allowed adverse drug effects to be assessed over time. Additionally, members of the project team had also worked with the THIN database before.

[7] IMI is a public–private partnership in the life sciences (according to IMI's website currently the world's biggest partnership of that kind), funded by the European Union and the pharmaceutical industry (IMI n.d.).

In this study, a random selection of seven drugs and associated adverse side effects (identified through THIN) were assigned to one of six assessors. Assessors were asked to manually review the link between the drugs and the adverse side effects using a specified questionnaire. vigiTrace, a framework for conducting exploratory analysis on EHRs, was also used to review the link between the drugs and side effects (Cederholm et al. 2015). They found that EHRs do contain safety signals, and detection of these requires both data analysis and human review (Cederholm et al. 2015).

2.3.5. What were the enablers and challenges of the study?

This study was enabled by the inclusion of the company Cegedim, the owners of the THIN database, in the team, allowing the study team to access the required data (Cederholm et al. 2015). In addition, the consortium of different organisations meant the project team had the required skills to conduct the study to a high quality. It was also considered helpful that one senior medical doctor involved was originally from the UK and therefore had an understanding of the UK primary care system.

The study faced challenges which slowed progress, such as frequent team member turnover and lack of funds, meaning much of the work had to be conducted remotely, which sometimes made progress difficult. There are also a number of factors which limit the generalisability of the study. For example, the study used only one EHR database which only included data from primary care settings and only covered the UK.

2.3.6. What were the safeguards employed to govern the use of the data?

To access the THIN database the study team required approval from the THIN Scientific Review Committee (UCL Institute of Epidemiology & Health Care n.d.).

2.3.7. What were the potential or realised benefits to patients and public health?

Introducing the use of EHRs alongside other methods to detect adverse side effects can allow for more accurate identification of potentially dangerous and fatal side effects, improving patients' quality of life (Cederholm et al. 2015). This topic continues to be explored by the Food and Drug Administration (FDA) in the United States (US) and in other European projects.

2.4. Case vignette 4: Use of non-randomised study data and observational data in network meta-analyses of RCTs to assess schizophrenia and in-stent restenosis treatment effects

This case vignette example illustrates the use of non-randomised study data and observational study data, in addition to RCT data, in assessing the effects of treatments for coronary in-stent restenosis and schizophrenia.

This case vignette is based on Efthimiou et al. 2017 and an interview.

2.4.1. Who was involved in the study?

This study was a collaboration between staff at the pharmaceutical company Eli Lilly, four universities (University of Ioannina School of Medicine, University of Ioannina, Technische Universität München and University of Bern) and two hospitals (Bern University Hospital and University Medical Center Utrecht) (Efthimiou et al. 2017). The team included a range of backgrounds including mathematicians, statisticians, a psychiatrist (for the schizophrenia study), a cardiologist (for the in-stent restenosis study), and an individual from Eli Lilly who oversaw the project and the safeguarding of the data.

2.4.2. What data were used?

The research for this case vignette focused on data related to the treatment of coronary in-stent restenosis and schizophrenia. Specifically, they used summary data from RCTs alongside summary data from non-randomised studies that assess drug effectiveness in the real world. For the coronary in-stent restenosis study they used data from 28 RCT papers (involving 5,914 patients) and six non-randomised study papers. For the schizophrenia study, they used evidence from 167 RCT papers (involving (36,871 patients) and non-randomised data from the observational study Schizophrenia Outpatient Health Outcome, which involved 8,873 patients from ten European countries. The data used in this study was provided by Eli Lilly (INT6).

2.4.3. What were the aims?

The aim of this study was to explore the use of non-randomised evidence, alongside RCT meta-analysis, to assess treatment effects, specifically for coronary in-stent restenosis and schizophrenia (Efthimiou et al. 2017), as well as to develop methodology for the reuse of real-world data in healthcare.

2.4.4. Why and how did they use the data?

Using only RCTs to assess the effects of treatment can become biased as they are only implemented in laboratory settings and have strict inclusion criteria. Including non-randomised data in assessment of treatments can reduce this bias and allow more precise data to be collected (Efthimiou et al. 2017). As the main purpose of the study was methodological, the researchers used the data to develop methods for the reuse of data.

2.4.5. What were the enablers and challenges of the study?

This study was enabled by access to the Schizophrenia Outpatient Health Outcome data, which was provided by Eli Lilly, who own the data. They were involved in the study from the grant writing stage.

Not having access to data can be a key barrier to a study of this kind. In the particular case of this study, another pharmaceutical company had agreed on providing data at the grant writing stage, but ultimately did not provide it. The publication used to produce this case vignette and the interviewee did not provide further challenges to the use of the data.

2.4.6. What were the safeguards employed to govern the use of the data?

The partners involved in this study had to sign a confidentiality agreement with Eli Lilly, who provided the data. They had to agree that all results would be anonymised as well as that data would be protected while being processed. A colleague from Eli Lilly was directly involved in the study to oversee the use of the data and ensure that processes are in line with the agreement. The provided data was also preprocessed and de-identified. All other data was already in the public domain and therefore free to access.

2.4.7. What are the potential or realised benefits to patients and public health?

This was a methodological study which showed that studies aiming to assess relative treatment effects, which are typically carried out based on RCTs, can benefit from also incorporating non-randomised studies. Such an approach has the potential to enhance the ability to understand which treatments are most effective, and therefore to prescribe the most appropriate treatment to patients.

2.5. Case vignette 5: Exploration into the feasibility of using Swedish EHRs and health registries to supplement or replace control populations in RCTs for asthma

This case vignette example illustrates a study exploring the feasibility of using EHR and health registry data from Sweden in clinical trial design. Specifically, the study investigated whether EHR data can supplement, or replace, the use of control populations in RCTs for asthma.

This case vignette is based on Franzén et al. 2016.

2.5.1. Who was involved in the study?

This study was a collaboration between staff at the pharmaceutical company AstraZeneca, three Swedish universities (Uppsala University, Karolinska Institutet and University Gothenburg) the health registry organisation Registry Center, and the statistics service provider Statisticon (Franzén et al. 2016).

2.5.2. What data were used?

The study used a linked dataset of EHRs from 36 primary care centres and one university hospital in Sweden, and a Swedish health register covering a total of 33,890 asthma patients and compared it against data from an RCT on severe asthma conducted by Astra Zeneca (Franzén et al. 2016).

2.5.3. What were the aims?

The study aimed to evaluate the feasibility of creating a 'real-world reference population' of asthma patients which could be used alongside control groups during long-term, RCT studies of asthma (Franzén et al. 2016).

2.5.4. Why and how did they use the data?

This study wanted to explore the possibility of using EHRs to supplement, or replace, a placebo group of patients in asthma trials as this could overcome the ethical challenge of randomising patients with severe asthma into placebo groups for long-term RCTs (Franzén et al. 2016).

The data from the EHRs and health registries were combined to create one data source. A control group was created out of this data and tested against the RCT data to see if it could be used instead of the control used in the RCT (Franzén et al. 2016). The study team found that the EHR data could supplement the control group for severe asthma (Franzén et al. 2016).

2.5.5. What were the enablers and challenges of the study?

The data used in this study lead to some limitations of the analysis. First, it was not clear whether inclusion and exclusion criteria were equivalent for the RCTs and the electronic health data which may have led to differences in the two populations. Additionally, they had to assume that treatment effects and outcome measures, such as asthma exacerbation, were measured in the same way between the RCT and EHR populations (Franzén et al. 2016).

2.5.6. What were the safeguards employed to govern the use of the data?

All data used in this study was de-identified (Franzén et al. 2016). The study protocol also received approval from a regional ethics committee in Sweden (Franzén et al. 2016).

2.5.7. What were the potential or realised benefits to patients and public health?

Using EHR and registry data can allow for collection of data on long-term safety and effects of asthma treatment. Additionally, using electronic health data may speed up clinical trials of treatments as RCTs often struggle to recruit patients for control groups. This may reduce the time it takes to conduct trials, potentially allowing effective treatments to reach patients quicker (Franzén et al. 2016).

2.6. Case vignette 6: Using paediatric pulmonary arterial hypertension health registries to improve understanding of the disease

This case vignette example illustrates how four paediatric pulmonary arterial hypertension disease registries were statistically analysed on an annual basis to improve understanding of the disease, particularly its development, progression and diagnosis. This approach may be more feasible than creating a new health registry.

This case vignette is based on Gliklich, Dreyer, and Leavy 2014.

2.6.1. Who was involved in the study?

This study was a collaboration between staff at the pharmaceutical company Actelion Pharmaceuticals and the EMA.

2.6.2. What data were used?

The study used data on paediatric pulmonary arterial hypertension (PAH) from four disease registries (a global registry and registries from the US, France and the Netherlands) (Gliklich, Dreyer, and Leavy 2014).

2.6.3. What were the aims?

The aim of this study was to identify a more effective method of evaluating the disease history, progression, development and treatment of paediatric PAH, a rare disease which is considered to be poorly described in typical paediatric populations (Gliklich, Dreyer, and Leavy 2014).

2.6.4. Why and how did they use the data?

With the advent of new therapies, paediatric PAH patients have a greater survival rate. There is therefore increased interest and need to investigate the disease development, progression and treatment experience. As paediatric PAH disease is rare, there was not already an existing data source which could be used for this.

Following the approval by EMA of Actelion Pharmaceutical's product for paediatric PAH, Actelion Pharmaceuticals, working with the EMA, developed a method to collect longitudinal data that could be used to monitor the population and outcomes related to their product. This method involved development of a common protocol to be used over a number of years to collect longitudinal data on paediatric PAH patients which could be used to examine data from patients in registries that already exist. As the registries use different data collection methods, each disease registry was analysed separately using the same common protocol. This analysis is conducted annually on each disease registry (Gliklich, Dreyer, and Leavy 2014).

2.6.5. What were the enablers and challenges of the study?

The study collected and analysed data yearly. In the first year they encountered challenges related to differences between the registries and how they analysed the data and interpreted the protocol. Time was

therefore spent on clarifying the requirements for subsequent years to ensure data is comparable (Gliklich, Dreyer, and Leavy 2014).

2.6.6. What were the safeguards employed to govern the use of the data?

After the EMA had reviewed and approved the study protocol, the four health registries also reviewed the protocol and agreed to participate. Additionally, the data sent to Actelion Pharmaceuticals was de-identified (Gliklich, Dreyer, and Leavy 2014).

2.6.7. What were the potential or realised benefits to patients and public health?

This study has exhibited the use of multiple registries to develop longitudinal cohorts for rare diseases, which can be used to improve understanding of rare diseases.

2.7. Case vignette 7: Using the Fabry Outcome Survey registry to explore the impacts of the disease on female reproductive and pregnancy outcomes

This case vignette example illustrates how statistical analysis of the Fabry Outcome Survey (FOS) registry was used in clinical research to explore the impact of this disease on female reproductive and pregnancy outcomes.

This case vignette is based on Hughes et al. 2018.

2.7.1. Who was involved in the study?

This study was a collaboration between staff at the biopharmaceutical company Shire, two universities (University College London and Castilla-La Mancha University), the statistical software developer Ctyel, and The National Institute on Population Medical Genetics (INAGEMP, which coordinates teams for disease research) (Hughes et al. 2018).

2.7.2. What data were used?

The study used the Fabry Outcome Survey (FOS) observational registry. This registry is funded by Shire and collects data on the safety and effectiveness of Fabry disease[8] treatment[9] and the disease history across 24 countries (Shire Outcome Surveys 2017).

2.7.3. What were the aims?

To use the FOS registry data to understand the impacts, if any, of Fabry disease on female reproductive and pregnancy outcomes (Hughes et al. 2018).

2.7.4. Why and how did they use the data?

The FOS registry provides an opportunity to understand the impacts of Fabry disease on patients. In this study, statistical analysis of the FOS registry was undertaken to determine the impacts of the disease on reproductive and pregnancy outcomes (Hughes et al. 2018).

2.7.5. What were the enablers and challenges of the study?

This study was enabled by the existence of the FOS registry. Additionally, the data in the registry was of high quality as the information is automatically checked for any logical inaccuracies when it is entered, and clarification questions can be sent to the data manager to fill in any missing information, although some data does remain missing (Hernberg-Ståhl 2006). On the other hand as the FOS registry is a disease specific registry, patients with more severe symptoms are more likely to be enrolled, and as there are not standardised assessments of Fabry disease the data quality can vary (Giugliani et al. 2016).

[8] Fabry disease is a rare, genetic condition causing a build-up of fat in cells causing various symptoms, such as hearing loss and pain in the hands and feet, and can lead to life-threatening kidney damage, heart attacks and strokes (NIH 2019).

[9] Enzyme replacement therapy (ERT) with agalsidase alfa.

2.7.6. What were the safeguards employed to govern the use of the data?

To ensure security and confidentiality of the data, the data is transferred to a central database managed by an external provider (Hernberg-Ståhl 2006).

2.7.7. What were the potential or realised benefits to patients and public health?

As Fabry disease is still not fully understood, analysis of the FOS registry could allow for a greater understanding of the disease and its impacts (Hughes et al. 2018).

2.8. Case vignette 8: Use of the THIN and Humedica EHR databases to explore any associations between BMI and an increased risk of non-alcoholic fatty liver disease

This case vignette example illustrates the use of two EHR databases, THIN and Humedica, in clinical research. The authors used the EHR data to explore the association between a higher body mass index (BMI) and greater risk of being diagnosed with non-alcoholic fatty liver disease.

This case vignette is based on Loomis et al. (2016).

2.8.1. Who was involved in the study?

This study is a collaboration between the pharmaceutical companies Pfizer and GlaxoSmithKline, the charity the British Heart Foundation, the University of Glasgow and the company Cegedim (who provided access to the THIN database) (Loomis et al. 2016).

2.8.2. What data were used?

The study used two EHRs: The Health Improvement Network (THIN) and the Humedica databases. The THIN database includes data on 12 million patients across the UK through GP surgeries and the Humedica database contains data on 25 million patients. For this study, data on 2.1 million patients were used across these two databases (Loomis et al. 2016).

2.8.3. What were the aims?

The aim of the study was to use the THIN and Humedica EHR databases to explore the relationship between a greater body mass index (BMI) and the risk of developing non-alcoholic fatty liver disease (NAFLD) (Loomis et al. 2016).

2.8.4. Why and how did they use the data?

This study aimed to understand population relationships between a disease and population characteristics; therefore, it required the use of EHRs (Loomis et al. 2016). Specifically, the study identified the association between BMI and risk of diagnosis with NAFLD (Loomis et al. 2016).

2.8.5. What were the enablers and challenges of the study?

This study was enabled by the existence of two EHRs containing patient characteristic data as well as NAFLD diagnosis (Loomis et al. 2016).

The data itself, however, has led to some limitations. For example, for some patients some characteristics (e.g. smoking status and BMI) were not included. These patients therefore had to be excluded from the study, reducing the population size. On the other hand, some patients may have had alcohol intake incorrectly recorded, meaning that patients who should have been excluded might not have been (Loomis et al. 2016). Additionally, the baseline prevalence of NAFLD in the two EHR databases are much smaller than in the real-world population, which is likely due to the under-diagnosis of NAFLD. The EHR databases also lack long-term follow up of patients, which can result in biased results (Loomis et al. 2016).

2.8.6. What were the safeguards employed to govern the use of the data?

Approval was provided to the study team from THIN and Humedica to analyse the data. This data was anonymised (Loomis et al. 2016).

2.8.7. What were the potential or realised benefits to patients and public health?

A greater awareness that a higher BMI is associated with a greater risk of being diagnosed with NAFLD is beneficial as NAFLD is often missed/overlooked when diagnosing patients (Loomis et al. 2016).

2.9. Case vignette 9: Using the Schizophrenia Outpatient Health Outcome dataset to explore the methods of identifying drug effect modifiers for schizophrenia

This case vignette example illustrates the use of the Schizophrenia Outpatient Health Outcome data, an observational dataset, to improve clinical trial design for schizophrenia through recruitment of more appropriate study participants. This was achieved by exploring potential methods to identify drug effect modifiers[10] for schizophrenia treatment.

This case vignette is based on Nordon et al. 2017.

2.9.1. Who was involved in the study?

This study was a collaboration between staff at the pharmaceutical company Eli Lilly, the market access services provider Analytica Laser, the research centre Bordeaux Population Health Research Centre, the University of Barcelona and the Farr Institute (Nordon et al. 2017).

2.9.2. What data were used?

The study used the observational dataset, the Schizophrenia Outpatient Health Outcome cohort. This includes data from 6,641 patients across ten European countries (Nordon et al. 2017).

2.9.3. What were the aims?

The aim of the study was to use the Schizophrenia Outpatient Health Outcome data to explore methods of identifying drug effect modifiers of schizophrenia to improve populations recruited for clinical trials (Nordon et al. 2017).

2.9.4. Why and how did they use the data?

The Schizophrenia Outpatient Health Outcome data were used for this study as it includes a wide range of data, including that related to the patient (their condition, prescribed treatments and outcome), but also data on the doctors (Nordon et al. 2017).

The study team carried out statistical analysis (mixed multivariable linear regression models) on the Schizophrenia Outpatient Health Outcome data to estimate the efficacy of antipsychotic drugs for schizophrenia, to explore the treatment similarities between patients treated at the same healthcare centre and to assess any changes in schizophrenia drug effects due to differences in the healthcare centres (Nordon et al. 2017).

[10] Drug effect modifiers are factors related to the patient, drug use and healthcare system which can influence the efficacy of a drug, such as patient age, drug dose or access to healthcare services (Miettinen 1974).

2.9.5. What were the enablers and challenges of the study?

The Schizophrenia Outpatient Health Outcome data were made available to the study team. An additional enabler was that the Schizophrenia Outpatient Health Outcome data included data from a large number of patients across multiple countries (Nordon et al. 2017). The publication used to produce this case vignette (Nordon et al. 2017) did not list challenges to the use of the data.

2.9.6. What were the safeguards employed to govern the use of the data?

The data used in this study was de-identified (Nordon et al. 2017).

2.9.7. What were the potential or realised benefits to patients and public health?

The results from this study help to identify the drug effect modifiers of schizophrenia treatment. This can help with the design of clinical trials which are accurate at identifying the effectiveness of treatments, and specifically ensuring the best trial population is used (Nordon et al. 2017).

2.10. Case vignette 10: Using EHRs to identify negative symptoms of schizophrenia and explore if these are associated with clinical outcomes

This case vignette example illustrates the use of South London and Maudsley NHS Foundation Trust EHR data for clinical research into schizophrenia. Specifically, the study aimed to identify the negative symptoms associated with schizophrenia within the general schizophrenic population and if there are any relationships between these symptoms and clinical outcomes.

This case vignette is based on Patel et al. 2015.

2.10.1. Who was involved in the study?

This study was a collaboration between staff at the pharmaceutical company Roche, two universities (King's College London and the University of Sheffield), and one NHS trust, the South London and Maudsley NHS Foundation Trust (who provided the EHR data) (Patel et al. 2015).

2.10.2. What data were used?

This study used EHR data from the South London and Maudsley NHS Foundation Trust's Biomedical Research Centre Case Register. This includes data of 1.2 million patients across south-east London receiving mental health care (Patel et al. 2015).

2.10.3. What were the aims?

The aim of the study was to use the EHR data to identify negative symptoms of schizophrenia and explore whether there is a relationship between these symptoms and clinical outcomes.

2.10.4. Why and how did they use the data?

Typically, the negative symptoms of schizophrenia are identified through studies of small population sizes, often with patients showing severe symptoms, and so cannot be generalised to the total schizophrenia population. This study wished to show that it was possible to use EHRs to identify negative symptoms of schizophrenia using a population more similar to the overall schizophrenic population, as this would allow collection of more generalisable results and thus more effective research into treatments (Patel et al. 2015).

To explore the data, the Clinical Record Interactive Search application was used and statistical analysis was conducted on the EHR data (Patel et al. 2015). This application was developed by the South London and Maudsley NHS Foundation Trust and allows authorised researchers to access de-identified information from the trust's EHR data (NIHR n.d.).

2.10.5. What were the enablers and challenges of the study?

The EHR data represents the overall clinical population of patients with schizophrenia and includes a large number of patient's data (Patel et al. 2015). The use of the Clinical Record Interactive Search application allows researchers to identify clinically relevant patterns more easily and reduces the time it

takes to conduct studies (NIHR n.d.). The publication used to produce this case vignette (Patel et al. 2015) did not list challenges to the use of the data.

2.10.6. What were the safeguards employed to govern the use of the data?

The EHR data is anonymised, is stored securely and can only be accessed by authorised researchers (NIHR n.d.).

2.10.7. What were the potential or realised benefits to patients and public health?

The use of the Clinical Record Interactive Search application allows research to be conducted faster, meaning schizophrenia treatment can be improved and thus patients can access treatment in a shorter space of time (NIHR n.d.).

2.11. Case vignette 11: Testing a data derivation tool to identify individuals with type II diabetes across four different types of health data sources

This case vignette example illustrates the use of four types of European data sources (primary care, record-linkage systems, hospital and biobank data) to test a data derivation method with the aim of improving studies conducted on data which span multiple countries and data sources. This study tested this data derivation method by identifying patients with type II diabetes across the four types of data sources.

This case vignette is based on Roberto et al. 2016 and an interview.

2.11.1. Who was involved in the study?

This study was a large collaboration between the pharmaceutical companies Pfizer, Janssen and GlaxoSmithKline; three healthcare centres (the Erasmus University Medical Center, the Boston Children's Hospital and Tartu University Hospital), the charity the British Heart Foundation, Cegedim, three universities (the University of Tartu, the University of Manchester and Universitat Pompeu Fabra), the Italian College of General Practitioners and Primary Care, Veneto's Research Centre for eHealth Innovation and the health data analysis collaboration Epidemiology and Observational Health Data Sciences and Informatics (Roberto et al. 2016).

2.11.2. What data were used?

Eight data sources were used for this study: three were primary care (the Health Search IMS Health LPD database, the Integrated Primary Care Information database and the THIN database), three were record linkage systems (obtained from Aarhus University Hospital, PHARMO and the Regional Health Authority of Tuscany), one was hospital data (obtained from Information System of Parc de Salut Mar Barcelona) and one was biobank data (obtained from Estonian Genome Center of University of Tartu). Overall, this included data from 12 million patients across six European countries (Roberto et al. 2016).

2.11.3. What were the aims?

The aim of this study was to test the use of a novel data derivation method to improve cross-country and cross-data source studies. In particular, this study aimed to use this data derivation method on a set of heterogeneous data sources to identify patients with type II diabetes.

2.11.4. Why and how did they use the data?

The study chose a range of different health data sources as they wished to test the data derivation method for its use across different types of health data (Roberto et al. 2016). Across the eight datasets, they conducted descriptive, cross-sectional and retrospective analyses. Case-finding algorithms were used to identify patients with type II diabetes within the data (Roberto et al. 2016).

2.11.5. What were the enablers and challenges of the study?

This study was part of the European Medical Information Framework (EMIF) project. As the eight data sources used for this study were included in the EMIF project, the authors were able to access the data (Roberto et al. 2016). In addition, the collaboration across the many partners was seen as beneficial as a range of thoughts and ideas were contributed to developing and conducting the project, helping to highlight the weaknesses as well as the strengths of their data. The size of the collaboration, both in terms of the number of different partners involved and the large number of work streams, also posed some challenges in terms of progressing the work.

As with all studies there are a number of limitations to the analysis. Namely, the algorithm used to identify patients with type II diabetes had not been validated and so relied on the judgement of experts which may have been subjective (Roberto et al. 2016). In addition, as the data used was originally collected for other purposes, it was difficult to apply it to this particular context and required assumptions to be placed on the data.

2.11.6. What were the safeguards employed to govern the use of the data?

Safeguarding of the data is not covered within the publication.

2.11.7. What were the potential or realised benefits to patients and public health?

This was a 'prototype' study to test the ability of creating a data derivation tool. Since this project, two other similar projects have been conducted through IMI and the tool has now been tested on chronic, acute and infectious diseases, generating evidence for its use across health settings.

This will allow for a greater understanding of how a drug works once it is on the market by a larger and more diverse populations than that of the clinical trial population, such as understanding why and how the drug is used and its safety and efficacy.

2.12. Case vignette 12: Using EHRs to obtain NICE reimbursement approval for a lung cancer treatment

The case vignette illustrates how Roche used FlatIron's EHR data to collect evidence for National Institute for Health and Care Excellence (NICE) reimbursement approval of Atezolizumab, a drug to treat lung cancer, after NICE had analysed RCT data and decided not to approve reimbursement.

This case vignette is based on two interviews.

2.12.1. Who was involved in the study?

The study was a collaboration between the pharmaceutical company Roche, and the cancer healthcare and technology services company FlatIron health[11] working with the UK National Institute for Health and Care Excellence (NICE).

2.12.2. What data were used?

The study used EHR data provided by FlatIron Health.

2.12.3. What were the aims?

Roche aimed to analyse real-world data provided by FlatIron Health to overturn the NICE decision to not approve reimbursement the drug Atezolizumab, which is used to treat lung cancer.

2.12.4. Why and how did they use the data?

NICE had analysed data from the RCT of another similar lung cancer treatment, Docetaxel. This RCT lasted for two years but NICE used a model based on this data to predict five-year survival rates. NICE came to the conclusion that the benefit it provided was not enough to cover reimbursement costs of the drug. Roche disagreed with this as they felt that the method NICE used underestimated the five-year survival rate. Roche then used the model used by NICE to test survival rates of Atezolizumab using real-world data (using Kaplan-Meier and log-logistic curve analysis) from FlatIron Health. This real-world data represented the same population as the NICE analysis to estimate the effects of the drug on a real cohort. The results from this were presented to NICE who changed their decision (NICE 2018).

2.12.5. What were the enablers and challenges of the study?

This study was enabled by the relationship between Roche and Flatiron Health and the skills held within these companies with relation to the use of EHR data, as Flatiron Health specialises in these analyses. The interviewees did not list challenges to the use of the data.

[11] FlatIron Health was acquired by Roche in 2018 and is part of the Roche group

2.12.6. What were the safeguards employed to govern the use of the data?

FlatIron follows privacy laws and restricts disclosure of confidential information within the Roche group (Flatiron n.d.).

2.12.7. What are the potential or realised benefits to patients and public health?

NICE now recommends use of Atezolizumab (NICE 2018), giving patients in the UK access to this treatment.

3. How is the pharmaceutical industry reusing health data?

This chapter focuses on understanding the specific types of health data that are (currently) reused by the pharmaceutical industry; how these health data are reused by the industry and what are they being used for; and the key reasons for the reuse of the data (i.e. why is the pharmaceutical industry using these data). Related to these core questions, we cover additional associated aspects of health data reuse by industry such as the extent of reuse (i.e. how much is the industry using health data), access to safeguarding of the data and the impacts (both positive and negative) of using the data. The analysis presented in this chapter is informed by the targeted literature review, information from the case vignettes and insights provided by the interviewees. As noted in Chapter 1, the evidence synthesised from the stakeholder interviews has been referenced using anonymised unique interview identifiers (INT1, INT2, etc.). We also use unique identifiers (CV1, CV2, CV3, etc.) to cite case vignettes in this chapter.

3.1. What types of health data are being reused by the pharmaceutical industry?

Key takeaways

- EHRs, health registry data and clinical trial data are the most frequently mentioned types of health data being used by the European pharmaceutical industry.
- EHR data are used particularly frequently as there are established databases to access these, e.g. the UK's CPRD.
- Other types of health data used by the European pharmaceutical industry include biobank data, prescribing and dispensing data, and claims data.
- More recently, less traditional types of real-world data are being used, e.g. social media data or data from wearables.

The literature review, case vignettes and stakeholder interviews demonstrate that a range of different types of health data are already being reused by the pharmaceutical industry. Multiple interviewees confirmed that the types of health data being used depends on factors such as the stage of the research, the methodology being used, the stage in the innovation pathway (i.e. from discovery to post-marketing), the type of the company, the location (e.g. Europe vs US), or more generally on whether a company can access the data, e.g. through buying data from vendors or collaborations with data owners (INT4, INT5, INT9, INT12, INT14, INT16). Across the case vignettes and interviews in particular, the use of EHRs was most frequently mentioned compared to other types of health data (CV2, CV3, CV5, CV8, CV10, CV11, CV12, INT4, INT5, INT8, INT9, INT10, INT14, INT15, INT16). Several interviewees mentioned the UK's CPRD (CV2, INT5, INT8, INT9, INT14, INT15, INT16), which was seen as one

of the most important EHR databases used by the pharmaceutical industry. The following reasons were cited for its prominence: it has been around for a long period of time; it is a rich source of data as it comprises records of several million patients; the processes to getting access to the data is standardised and straightforward; and it is open to any kind of organisation, including pharmaceutical companies (CV2, INT5, INT8, INT16). One interviewee elaborated, however, that databases such as CPRD may not be easily accessible by pharmaceutical companies alone, but it can be made easier if they are collaborating with other stakeholders (INT16). The use of EHR data was exemplified in several of the case vignettes included in the analysis (CV2, CV3, CV5, CV8, CV10, CV11, CV12). For example, in a collaboration effort between a pharmaceutical company, a health centre and two universities, CPRD data were used in post-release research to investigate how glucose-lowering drugs have been prescribed; the aim was to assess bias in estimates of effectiveness made from real-world data for these drugs in this setting (CV2). In other case vignettes, EHR data from The Health Improvement Network (THIN) were used for drug safety assessments (CV3), clinical research (CV8) and – in combination with other data types – to test a data derivation method with the aim of improving studies conducted on data which span multiple countries and data sources (CV11).

Interviewees also frequently referred to the secondary use of health registry data (Galson and Simon 2016; CV5, CV6, CV7, INT4, INT8, INT9, INT10, INT14, INT15, INT16) and clinical trial data (CV1, CV4, CV9, INT5). Further types of health data currently being used by industry include biobank data and other types of genetic data (Cole and Towse 2018; CV11, INT10, INT15), prescribing and dispensing data (INT4), and claims data (Cole and Towse 2018; Galson and Simon 2016; INT5, INT8, INT10, INT15). According to one interviewee, claims data are more common in the US, while countries like the UK do not have large claims databases available for research (INT5). More recently, less traditional types of real-world data are being used by some pharmaceutical companies, including social media data (Galson and Simon 2016; INT9, INT14), data from wearables (Cole and Towse 2018; Galson and Simon 2016; INT10) and other patient-generated data, such as data submitted by patients on online research networks such as the US-based PatientsLikeMe (Cole and Towse 2018; Khosla et al. 2018) or patient-reported outcomes (Cole and Towse 2018; Galson and Simon 2016). The literature also refers to the use of administrative data, prospective observational data, pharmacy data, patient pathway data, health survey data, cost study data, surveillance data, mobile device data, data form mortality databases, consumer data, and primary and secondary care data, although these were not mentioned by the interviewees (Cole and Towse 2018; Galson and Simon 2016).

Some of the case vignettes showed that different types of health data are combined in research and development. For example, a pharmaceutical company, in collaboration with European universities and a hospital, used published clinical trial data and data from a non-randomised trial for an assessment of the effects of treatments for coronary in-stent restenosis and schizophrenia (CV4). Similarly, a research team involving a pharmaceutical company, a health registry organisation, several universities and a statistics service provider combined EHR and health registry data to explore the feasibility of using these types of data in clinical trial design (CV5).

3.2. Why is the pharmaceutical industry reusing health data?

Key takeaways:

- Real-world data enable better insights into the real world than an artificial setting (such as clinical trials) could.
- Secondary health data analyses lead to more efficiency and help reduce costs for both industry and health systems.
- The growing availability of health data is contributing to their increasing use.
- Secondary analyses can allow for control groups to be constructed from data that already exists, rather than needing to enrol patients.

The literature, case vignettes and interviewees cited numerous reasons as to why the pharmaceutical industry is reusing health data. The ability of real-world data to provide better analyses and better insights into the real world was most frequently mentioned as a driver of using such data (Galson and Simon 2016; Khosla et al. 2018; Singh et al. 2018; CV3, CV4, CV10, INT1, INT2, INT4, INT6, INT8, INT9, INT10, INT12, INT13, INT14, INT15, INT16). Real-world data analyses provide a 'better' picture of a drug than the relatively constrained nature of clinical trials as the difficulty with clinical trials data is that it is hard to generalise these data to the wider population (Galson and Simon 2016; INT2, INT4, INT8, INT9, INT10, INT12). Moreover, clinical trials data can potentially result in bias due to the restricted settings they are implemented in (e.g. strict exclusion criteria such as excluding people who smoke or drink). While such criteria are essential for clinical trials, these restrictions do not apply to 'real-world' settings, where drugs are prescribed to patients who would not have qualified for the clinical trial (CV4, INT6, INT10). Using real-world data, in particular in conjunction with clinical trials, facilitates the pharmaceutical industry to understand the wider impact of a drug in the real world, and for the results of RCTs to be generalised across a wider population. In addition, adding a real-world data component to a clinical trial setting provides better validity than only clinical trials themselves. For example, if EHR data are used in this scenario, these could help show that a treatment is reliable as it has been tested against data from the real world (INT14). As highlighted in one of the case vignettes (CV2), for example, some analyses are not possible without real-world data. In post-release research on prescriptions of glucose-lowering drugs, researchers used EHR data to study whether the characteristics of patients being prescribed diabetes treatments differ for different treatments (i.e. whether there is channelling bias), and in particular whether those patients being prescribed new diabetes treatments are different from those being prescribed two existing drug classes (Ankarfeldt et al. 2017; CV2). Investigating such channelling bias is not possible without real-world data (CV2).

Interviewees also differentiated between the types of data being used. Health data collected for medical reasons – EHR data, in particular – were seen as more reliable and often more accurate than other types of data as they have typically been collected across the whole patient population; they can therefore be better generalised to the patient population (CV3, CV10). Claims data, by contrast, were seen as less reliable as they are collected for purposes other than health reasons (CV3).

Efficiency was often highlighted as a reason for reusing health data (INT2, INT4, INT9, INT12, INT14, INT15). Using data that are already readily available enables better use of resources. Related to this, it was noted that it can be less expensive to conduct a secondary analysis of data than running a new clinical trial

(INT5, INT9, INT14). One interviewee argued that there are not only potential cost savings for the pharmaceutical industry when existing health data are used, but the reuse could also reduce public money spent on research (INT12).

The use of real-world data is also considered as a way to contribute to improving the access to treatment. Some interviewees noted that certain types of analyses would not be possible without real-world data (CV2, CV8, INT12, INT15). For example, real-world data enables a greater use of precision medicine; according to one interviewee, precision medicine cannot be done 'without proper real-world data and without proper understanding of how patients go through their disease' (INT12). Interviewees from industry also noted that real-world data help target treatments to populations that are most likely to benefit from them as they provide detailed information on what types of patients need what types of treatments (INT4, INT5, INT15). The pharmaceutical industry is also using health data to expand treatments to other indications (Khosla et al. 2018; INT9). Moreover, some interviewees noted that their companies are using health data because this may enable quicker access to the market if a full clinical trial is not needed (INT9, INT14).

Health data are also often used because of convenience: some data are readily available and industry has access to them (Galson and Simon 2016; CV2, CV3, CV4, CV5, CV7, CV9, CV11). Some companies in the pharmaceutical sector conduct secondary analyses because they can place less burden on patients than traditional clinical trials (CV5, INT10, INT12). For example, one interviewee highlighted that some companies use real-world data to develop a comparator arm[12] in clinical trials (instead of recruiting new patients) on conditions such as cancer (INT10). In another example, EHRs were tested as a replacement for a placebo group of patients in asthma trials to address the potential ethical issues which might be associated with randomising patients with severe asthma into placebo groups for long-term RCTs (CV5).

An interviewee also felt that the use of health data has a positive impact on collaborations. As access to health data was seen to be easier and less expensive if these are provided by partners, the use of health data requires the pharmaceutical industry to actively develop and nurture collaborations with entities like universities and hospitals (INT15). Engaging in such collaborations and using data from the real world contributes, in turn, to creating credibility. According to an interviewee, 'people would be suspicious if you create everything yourself', whereas 'having an externalised relationship and showing that we are actually using data from the real world creates credibility' (INT15).

[12] Comparator arms of a clinical trial are groups or subgroups of people that, for example, receive an already existing treatment or intervention, placebo or no treatment/intervention at all. These groups of people are compared to a group of people that receive the new treatment/intervention to assess the effectiveness of this new treatment/intervention (ClinicalTrials.gov 2019).

3.3. What is the pharmaceutical industry reusing health data for?

Key takeaways

- At the discovery and drug development stage, real-world data are used:
  - To identify diseases or indications of a significant burden to a wider population.
  - To better understand a disease, e.g. the impacts of a disease on the wider health and well-being of patients, risk factors associated with a disease or disease progression.
  - To understand the prevalence of a disease or condition.
  - To provide new insights into disease associations or comorbidities and therefore to target new populations and indications for future research.
  - To develop targeted and personalised therapies and drugs.
  - To develop new analytical methods.
- At the clinical research stage, real-world data are used:
  - To inform clinical trial design, e.g. to improve the study population selection for clinical trials, to predict the number of potential patients or to assess the efficacy of a new drug.
  - To create new approaches to patient stratification.
  - In feasibility studies.
  - Alongside or instead of control groups for trials to reduce the need to enrol patients as controls.
- At the marketing and sales stage, health data are used:
  - For medicine authorisation and regulatory purposes.
  - To support market access discussions, e.g. to conduct health technology assessments (HTAs), identify how competitive drugs are used on the market and to support pricing discussion.
  - To conduct cost-effectiveness analyses.
- At the post-authorisation stage, health data are used:
  - To support pharmacovigilance, i.e. to identify safety issues and adverse reactions.
  - For pharmacoepidemiology, i.e. to understand treatment effects across patient populations, to identify patient groups resistant to drugs, as well as to get insights into patient adherence.
  - To add to the medical evidence base and inform changes in practice guidelines.
  - To support effectiveness comparisons between new drugs and existing drugs.
  - To inform drug repurposing, i.e. the identification of diseases and conditions that could be treated with an existing drug.

The European pharmaceutical industry uses health data across all stages of the research and innovation (R&I) pathway, i.e. from the discovery and development stage,[13] through the clinical research stage, to the marketing, sales and post-authorisation stage (Miani et al. 2014). Below, we describe how health data are used by industry in each stage of the R&I pathway.

[13] The discovery stage of the pharmaceutical treatment development process is the stage at which diseases/conditions are being researched to gain new insights into a disease/condition as well as to identify targets for new drugs. It is also the stage at which researchers search for chemical compounds (which could become the new drug) that may be able to act on this target. These compounds are being tested to determine if they are able to act on this target. Once the search for compounds has been narrowed down to a few compounds, these compounds will be taken to the development stage. At this stage, compounds are being tested regarding their effectiveness and safety, and may be taken forward to the clinical research stage (J. P. Hughes et al. 2011).

3.3.1. Discovery

There are a range of use cases of health data at the discovery stage, although interviewees felt that real-world data are currently less used at this stage than at other stages in the innovation pathway (INT2, INT4, INT5, INT8, INT9, INT10, INT12).

Health data are being used to identify diseases or indications of a significant burden to a wider population (Khosla et al. 2018) and to better understand a disease (CV7, CV8, CV10, INT15). In two case vignettes (CV7, CV10), for example, registry data and EHR data were used to study the impacts of a disease and its symptoms on the wider health and wellbeing of patients, and in one case vignette (CV8) researchers used EHR data to explore whether people with a high BMI are at a higher risk of developing NAFLD. The pharmaceutical industry also uses real-world data to understand disease progression, i.e. how a disease develops and changes over time (INT2, INT5, INT8, INT9, INT10, INT14). In a study on paediatric PAH, for example, a pharmaceutical company conducted research with data from four disease registries to understand how the disease develops and progresses over time (CV6). An interviewee noted that the discovery stage is the most common stage where genetic data are currently used (INT15).

Two interviewees from pharmaceutical companies mentioned that they use real-world data to understand the prevalence of a disease or a condition (INT10, INT12). As highlighted by an interviewee, disease registry data are most commonly used to assess prevalence, whereas EHR data are less suitable for such analyses, particularly if the disease or condition of interest is rare; in such cases, there may not be sufficient amount of data to carry out meaningful analyses (INT10). Registry data allows the collection of cohort data from patients over a long period of time and registry data better links data from primary, secondary and tertiary healthcare (INT10).

In addition, real-world data are used to provide new insights into disease associations or comorbidities, and thus can be useful for targeting new populations and indications for future research (Singh et al. 2018). For example, in one of the case vignettes, EHR data enabled the research team to find out that a greater BMI is associated with the risk of developing NAFLD (CV8). Moreover, several interviewees and studies reported in the literature explained that real-world data are used to develop targeted and personalised therapies and drugs (European Medicines Agency 2016; Singh et al. 2018; INT1, INT3, INT8, INT9, INT10). For example, large sets of genomics data can provide better insights into a disease and how it unfolds in different types of people, and therefore enable stratification – i.e. the identification of differences in treatment effects in different groups of patients (Roitmann, Eriksson, and Brunak 2014) – and also enable the development of treatments personalised to a patient or patient group (European Medicines Agency 2016). In a number of European countries, there has been significant effort in establishing large genome projects and systems. For example, the French Plan for Genomic Medicine 2025, a ten-year plan, aims to fully integrate genomic medicine into healthcare across the country as well as to establish a genomic medicine sector nationally (European Society for Medical Oncology 2018). Similarly, the Genomics England 100,000 Genomes Project aims to create a national genome research platform to improve the development of personalised medicine (European Society for Medical Oncology 2018).

As evidenced in several of the case vignettes in this report, real-world data are also being used to develop new analytical methods. For example, in a study conducted by a pharmaceutical company and several

European universities, researchers used published RCT data and non-randomised study papers to develop new methodological approaches to secondary analyses of health data (CV4). Similarly, another pharmaceutical company aimed to identify a more effective method of evaluating the disease history, progression, development and treatment of paediatric PAH (CV6). In two other case vignettes, pharmaceutical industry researchers (CV3) and a team of industry and academic researchers (CV5) explored whether real-world data – specifically EHR and registry data – are a valid tool for pharmaceutical research and whether such data could be used alongside, or instead of, control groups for trials to overcome the ethical challenges of having patients with severe diseases or conditions in the placebo groups for RCTs, which can be critical for their cure (CV5). A consortium of pharmaceutical companies, hospitals, health centres and academic researchers across Europe also tested the use of a new tool to improve the analysis of data spanning different countries and datasets (CV11).

3.3.2. Clinical research

Real-world data have already become a key element of clinical trials in the European pharmaceutical industry (European Commission 2016; Khosla et al. 2018; CV9, INT1, INT2, INT4, INT5, INT8, INT9, INT12, INT16). Such health data are increasingly being used to design clinical trials, e.g. in addition to or instead of the control arms to calculate the control response (INT2, INT4, INT9, INT10, INT15); they help improve the study population selection for clinical trials as they support the identification of the patients that are most likely to benefit from a new treatment (CV6, INT5, INT8, INT15, INT16), and researchers can then follow these patients over time (INT8, INT15); they can be used to predict the number of potential patients (INT12); and they can be used to assess the efficacy of a new drug, especially in populations which were small in the clinical trials (INT2). Interviewees felt that real-world data are particularly useful for the protocol development as they can give timely feedback at this stage (INT9, INT12, INT15). For example, EHR data can be used to optimise the inclusion/exclusion criteria before a clinical trial starts; such timely feedback is critical as any amendment of a protocol at a later stage can result in delays of several months as well as incur significant additional costs (Khosla et al. 2018; INT8, INT9, INT12). An interviewee gave a hypothetical example of a clinical trial on rheumatoid arthritis: according to the protocol, the researchers would need 500 patients for their trial, but at the recruitment stage they could only find 476 participants. If something like this is already known at design stage, any adaptations to the protocol could already be made at an early and less critical stage (INT8).

Studies reported in the literature also referred to the use of real-world data to create patient stratification (European Medicines Agency 2016; Singh et al. 2018), i.e. to divide the targeted patient group into subgroups to identify differences in treatment effects (as well as other factors such as adverse effects) across different groups (Roitmann, Eriksson, and Brunak 2014). Singh et al. (2018) reviewed four studies where real-world data were used to develop novel approaches to patient stratification;. For instance, in one study patient records were used to identify individual risk factors after a knee arthroplasty (Singh et al. 2018).

As highlighted in Section 3.2, real-world data are also used alongside control groups for trials or instead of control groups as this would help overcome ethical challenges of having patients with severe diseases such as cancer in the placebo groups (INT10, INT12, CV5).

The literature also refers to other types of studies where health data are used, e.g. in feasibility studies (European Commission 2016).

3.3.3. Marketing authorisation and market access

Real-world data are sometimes used for medicine authorisation and regulatory purposes (European Commission 2016; European Medicines Agency 2016; INT1, INT2, INT5, INT8, INT9). Some interviewees felt, however, that European regulations tend to lag behind, for example, the US in terms of acceptance and in that they do not provide full guidance on the use of and requirements of real-world data (INT2, INT4, INT10, INT14). Nevertheless, European and international regulators, as well as other healthcare decision makers, are increasingly requesting the pharmaceutical industry to submit real-world data. For example, there has been an increase in regulator requests of 'post-marketing commitment' studies using real-world data as a condition of regulatory approval (Khosla et al. 2018). Some interviewees were convinced that the use of health data could improve regulatory submissions as they would provide 'better' evidence than clinical trial data alone (INT4, INT8, INT14, INT16).

Real-world data are used to support pharmaceutical market access discussions, for example to conduct health technology assessments (HTAs), to identify how competitor drugs are used on the market and to carry out analyses to support pricing discussions (Galson and Simon 2016; INT4, INT10, INT12, INT15). Real-world data are also key for the pharmaceutical industry to be able to better identify the number of patients the drug could benefit, and to conduct cost-effectiveness analyses. If a company is able to show that a drug has 'good' real-world outcomes, this can help ensure that the company gets reimbursement contracts (Ali et al. 2017; Khosla et al. 2018; 2017; INT8, INT10, INT14). An interviewee explained that it is helpful for the pharmaceutical industry to have country-specific health data to be able to support pricing negotiations at national level – given regional differences in healthcare systems, having such targeted information supports industry in their negotiations (INT10). In one of the case vignettes included in this report (CV12), a pharmaceutical company used real-world data to successfully appeal a negative reimbursement decision. The company developed a new lung cancer treatment, but the UK's NICE concluded that the benefit it provided was not enough to cover reimbursement costs of the drug. The pharmaceutical company disagreed with the decision as they thought that the model NICE used underestimated the five-year survival rate. The company then used this model to test survival rates of their drug using EHR data. They were able to demonstrate that the model NICE used was incorrect and the decision was adjusted accordingly (CV12).

3.3.4. Post-authorisation

Evidence suggests that real-world data are particularly useful for the pharmaceutical industry's work in pharmacovigilance (i.e. to identify safety issues and adverse reactions that were not apparent in the clinical trial). The analysis of real-world data enables a more reliable evidence base to be established to highlight the benefits and risks of a drug (Cole and Towse 2018; Galson and Simon 2016; CV2, INT2, INT4, INT5, INT8, INT9, INT15, INT16). Some regulators have also recognised the use of real-world data for pharmacovigilance in light of several treatments being withdrawn due to safety issues after approval and introduction to the market (Khosla et al. 2018).

In post-drug development research, real-world data are used for pharmacoepidemiology (i.e. to understand treatment effects across patient populations) to identify patient groups resistant to drugs as well as to get insights into patient adherence (Cole and Towse 2018; European Commission 2016; Singh et al. 2018; CV2, CV4, INT15). One interviewee discussed the importance of using real-world data to generate and add to the available medical evidence base. Specifically, health data can be used to fill the gaps in clinical knowledge and ensure that treatments are supported by evidence to allow for a change in practice guidelines to prescribe them (INT16).

Once a drug is on the market, real-world data can also support effectiveness comparisons between the new drug and existing drugs (Bate, Reynolds, and Caubel 2018; Khosla et al. 2018; Singh et al. 2018; INT10, INT14, INT16). An interviewee felt that the use of real-world data for such comparative purposes will likely increase in the future (INT5).

Real-world data are further used to inform drug repurposing, i.e. to identify diseases and conditions that could be treated with an existing drug (Khosla et al. 2018; Pushpakom et al. 2019).

3.4. To what extent is the pharmaceutical industry reusing health data?

Key takeaways

- The use of health data by the pharmaceutical industry has become more commonplace over the past few years. There has been a particular increase in its use at the discovery and drug development stages whereas previously it was mainly used at the market entrance and post-authorisation stages.
- The scale of real-world data use differs by the type of pharmaceutical company. European companies developing new drugs are using real-world data more extensively than, for example, manufacturers of generic drugs.
- In general, larger pharmaceutical companies tend to use real-world data more often than smaller ones.

The interviewees and the literature did not provide definitive evidence to indicate the scale of health data reuse by the pharmaceutical industry. However, some interviewees felt that there has been an increase in the secondary use of health data over the past few years (INT5, INT9, INT14). Furthermore, the extent of health data reuse within the pharmaceutical sector could be expected to increase in the future as their value is understood better and there is increasingly more interest from regulators (INT4, INT5, INT8, INT9, INT12, INT14). Moreover, the use of real-world data in predictive modelling and to develop personalised medicine as well as the use of biobank data are expected to increase in the near future (INT9). While previously, secondary analysis was mainly done during the launch of a drug or at post-marketing stages (Khosla et al. 2018), it is becoming more common in the discovery and drug development stages (Khosla et al. 2018; INT5). Another interviewee explained that while a few years ago health data were only used in a small fraction of projects, they expected that 'in one or two years maybe three quarters of non-interventional studies will have secondary data' (INT14).

A regulatory authority representative felt that especially at the discovery and drug development stages, the use of real-world data differs by the type of pharmaceutical company. Companies developing new drugs are already using health data extensively, while others, such as manufacturers of generic drugs, use them less often (INT4). Many larger pharmaceutical companies have set up their own real-world data teams over the past few years, while smaller companies will find it more challenging (for example, in terms of

available resources) to have teams dedicated to real-world data (INT5). Another interviewee thought that different-sized companies also have dissimilar priorities, and this may partly explain differences in terms of scale. Larger pharmaceutical companies are mainly interested in predicting responses to treatment whereas small and medium-sized enterprises (SMEs) are more interested in finding different outcomes that better reflect a disease or a response (INT9).

Despite several interviewees highlighting the rapid increase in the use of health data by the pharmaceutical industry (INT5, INT9, INT14), an interviewee also pointed out that much of what the industry has done so far (and is currently doing) is not in the public domain (INT1). Possible reasons for this include: protection of company interests; analyses used for regulatory purposes may not be able to be published until the regulatory process is completed; and there is not necessarily a need to publish exploratory and discovery research (INT1, INT3).

3.5. How does the pharmaceutical industry access health data and how are these data safeguarded?

Key takeaways

- Industry often use collaborations with other organisations (e.g. other pharmaceutical companies, healthcare providers, universities and research organisations, private organisations to regulators, and policymaking and arm's-length bodies) to get access to health data.
- Industry uses publicly available datasets as well as buys health data from vendors.
- The ease of accessing health data depends on the type of health data and the nature of the study they are intended for, as well as other contextual factors such as country.
- There is a spectrum of governance arrangements for access to health data, from publicly available (anonymous) data sets to completely restricted access.
- Access to health data is a particular challenge when a study requires multiple datasets, as this requires research teams to approach each dataset owner individually regarding access.
- There are several ways how the use of health data by industry is safeguarded:
    - Safeguards set up by the data owner.
    - Internal governance and control policies.
    - Restricted access to data within an organisation.
    - Ethical approvals.
    - Confidentiality agreements.
    - Privacy legislation.
    - Using the same security standards as used for clinical trials.
    - Policies restricting access such that analyses can only be carried out by the data owner or by specific organisations that have oversight of the data.

Several interviewees and case vignettes highlighted that collaborations among different stakeholder groups are common when it comes to reusing health data, with stakeholders ranging from other pharmaceutical companies, healthcare providers, universities and research organisations, private organisations, regulators, and policymaking and arm's-length bodies (Khosla et al. 2018; CV1, CV2, CV3, CV4, CV5, CV7, CV8, CV9, CV10, CV11, CV12, INT3, INT5, INT8, INT9, INT12, INT15, INT16). According to Khosla et al. (2018) and some interviewees (INT5, INT8, INT15), collaborations are critical for industry to get access to health data. While some health data are publicly available (INT9) or can be bought from vendors (INT2, INT5, INT8, INT12, INT14, INT15, INT16), these data are sometimes limited in

terms of the population covered or a part of the healthcare system (e.g. primary care) (INT5, INT12). For example, publicly or commercially available health data are often from countries outside Europe (e.g. the US) and it can be more difficult to access European health data due to stricter privacy laws (INT5, INT12). Additionally, European health data are often limited to primary care EHR and claims data; however, these EHR data tend not to include hospital data (INT5, INT12). Large EHR datasets such as the CPRD also include data solely from primary care settings, which limits the 'richness' of the data for some diseases or conditions that are typically treated in non-primary care settings.

An interviewee noted that dataset owners are linking their primary care data to hospital data more frequently, although the hospital data used is often limited to specific regions (e.g. London, northern England or Scotland) and not a countrywide effort (INT5). Cross-stakeholder as well as cross-country collaborations help overcome access issues. According to an interviewee, their company prefers collaborating with hospitals where clinicians collect the data, undertake collaborative research and analyses with their industry partner(s), and in turn the industry partners help the clinicians acquire better analytical skills (INT8).

As evidenced in two of the case vignettes (CV4, CV11), pharmaceutical representatives are sometimes not directly involved in analyses in such collaborations, but rather oversee or manage the project and safeguard the health data, while the academic researchers or clinicians conduct the analyses (CV4, CV11). For example, in a study assessing the effects of treatments for coronary in-stent restenosis and schizophrenia – which was funded by a joint undertaking between the European Union (EU) and EFPIA (Efthimiou et al. 2017) – a pharmaceutical company provided the data for the research, and all other project partners had to sign a confidentiality agreement with the company. This agreement set out that all results would be anonymised and that data would be protected throughout while being processed. A representative of the pharmaceutical company oversaw the use of the data to ensure that processes were in line with the agreement (CV4).

According to some interviews, how health data are accessed depends on the type of health data being used, the nature of the study and the country context (INT14, INT15). For some datasets, those wanting to use them must apply for permission to gain access (CV1, CV2, CV10). Moreover, in some cases ethical approval is required, often from an independent committee or body (CV2, CV3, CV5, CV6, CV8, INT9), or a confidentiality agreement to access the health data needs to be signed (CV4).

Access to health data is a particular challenge when a study requires multiple datasets, as this requires research teams to approach each dataset owner individually regarding access (INT9). Another related challenge is data ownership: several data owners do not allow pharmaceutical companies to access their health data (INT9, INT14). An interviewee felt that there is a need for a proactive 'conversations' between policymakers, academia and industry about how governments and industry could get access to health data produced by academia and vice versa (INT14); a similar point was made in a workshop of the UK Academy of Medical Sciences and the Association of the British Pharmaceutical Industry (ABPI) (The Academy of Medical Sciences 2016).

Apart from ethical approvals and confidentiality agreements (CV2, CV3, CV4, CV5, CV6, CV8, INT5, INT9), there are a range of ways to safeguard the use of and access to health data, and it is usually a combination of several safeguards (INT15). As several interviewees pointed out, as is the case with every

organisation, industry has to follow European privacy legislation (INT4, INT8, INT10, INT15), and industry is trying to apply the same security standards they also use for clinical trials and have internal governance and control policies in place (INT8, INT15). Such internal governance and control policies often specify that access to the data is limited to a few people. More generally, safeguards are usually set up by the data owner (INT5), which could include (for example) de-identification or anonymisation[14] of the data prior to sharing them with the buyer (CV1, CV2, CV3, CV4, CV5, CV6, CV8, CV9, CV10, INT9, INT1, INT12, INT14).

In cases of collaborations across organisations, health data are often owned by and also confined to one organisation, or an individual of this organisation oversees the use of the data, and ensures that they are used ethically and in line with access agreements (CV3, CV4, CV12). In some cases, health data are transferred to and stored with an external database manager to ensure security and confidentiality of the data; or analyses are done by external bodies (e.g. the vendors or academic partners), which means that some companies conducting work with health data never have direct access to these, but only receive reports detailing the results once the analyses have been done (CV12, INT14, INT15).

[14] De-identification means 'removing or obscuring any personally identifiable information from individual records in a way that minimises the risk of unintended disclosure of the identity of individuals and information about them' (Nelson 2015, 12). It is therefore possible to *de*-de-identify data again, whereas anonymisation 'refers to the process of data de-identification that produces data where individual records cannot be linked back to an original as they do not include the required translation variables to do so' (Nelson 2015, 12).

3.6. What are the potential impacts of the secondary use of health data?

Key takeaways

- The literature and interviewees reported more often on the potential positive impacts of the use of health data than on negative impacts. Several interviewees felt that there were no negative impacts at all.
- Identified potential positive impacts include:
    - Improved treatments for patients.
    - Accelerated research and drug development, and therefore also accelerated treatment access.
    - Better understanding of and improvement of treatment efficacy.
    - Improved pharmacoepidemiology and pharmacovigilance.
    - Most appropriate patient populations can be recruited.
    - If research and development processes are accelerated, industry can get their products earlier on the market.
    - Secondary analyses can support pricing discussions, and thus help companies get reimbursement contracts for their drugs.
- Interviewees did not refer to actual examples of negative impacts, but only to potential ones (note that this does not indicate that there are no negative impacts associated with the reuse of health data):
    - Negative impacts of reuse of health data could occur if pharmaceutical companies broke personal data laws or if they used the data for purposes other than the original purposes. This could lead to less trust and restricted access to health data.
    - Negative impacts of reuse of health data could occur if pharmaceutical companies conducted poor-quality analyses, e.g. if they 'fished' for the results they want to find.
    - If health data have low quality, this could lead to poor research results and variable evidence.

Across the various sources of evidence analysed for this study (literature, interviews and case vignettes), (potential and realised) positive impacts of the use of health data were more often reported than any (potential and realised) negative impacts.

In terms of positive impacts, both societal impacts and benefits for the pharmaceutical industry were mentioned. Interviewees frequently emphasised that secondary analyses can lead to improved treatments for patients. As the use of real-world data can lead to a better understanding of a disease as well as provide insights into the effects of a treatment across different patient groups, it may enable physicians to more effectively select the right treatment for their patients (CV2, CV4, CV5, CV10, CV11, CV12, INT12, INT15). The use of health data could also accelerate research and drug development processes – for example, if a full clinical trial is not needed – and thus should eventually also accelerate treatment access (CV5, CV10, INT9, INT14). Moreover, if there is a better understanding of a drug and how it works – also across different patient groups and for what types of diseases and conditions – it could be made available to more patients (CV11, INT9) and the right treatment could reach the right patients quicker and more often (INT4, INT5, INT15). Proving the efficacy of a drug can support reimbursement decisions, and if a drug's costs are getting reimbursed, the uptake by healthcare organisations may be higher, allowing patients to access potentially better treatments (CV12, INT2, INT8, INT10, INT12) as well as to enable industry to get their product more swiftly on the market. A better understanding of the risk factors of diseases can also improve the accuracy of diagnosis in those that are frequently missed or misdiagnosed (CV8).

A further positive impact for the research community, and eventually for patients, is that secondary analyses of health data help understanding of treatment efficacy and ensures that efficacy is not only considered in the constrained environment of a clinical trial, but also in the real world. For example, as highlighted in the case vignettes, secondary analysis can demonstrate that a treatment is as effective as the clinical trial results suggested, as well as allow for a better understanding of how a drug works after it has been released in the market (CV2, CV11, INT2, INT15). In addition, health data such as EHRs and registry data can give the pharmaceutical industry insights into the long-term efficacy and safety of a treatment (CV5, INT15).

Real-world data can improve pharmacoepidemiology and pharmacovigilance, help detect adverse drug events more rapidly and thus contribute to drug safety (European Commission 2016; Singh et al. 2018; CV2, CV3, CV5, CV11, INT15). This would allow health professionals to be better able to provide the most appropriate treatment (CV2). More accurate identification of adverse side effects also reduces the negative impact on patients that could result from side effects, and therefore improve the quality of life (CV3).

Secondary analyses also bring several (potential) benefits to pharmaceutical companies: they enable better designs of clinical trials as the most appropriate patient populations can be recruited (CV9, INT8). Using real-world data in clinical trials, especially if these are datasets containing information about a large number of people, allows for more accurate and more rapid analyses at a lower cost (CV5, CV8, CV10). Accelerating research processes also enables pharmaceutical companies to get their products earlier on the market (CV5, CV10). Interviewees also suggested that demonstrating the efficacy can also support the pharmaceutical industry in pricing discussions and help companies get reimbursement contracts for their drugs (CV12, INT2, INT8, INT10, INT12), and thus be a further (monetary) benefit to the pharmaceutical industry.

As noted above, positive impacts of using health data were more often mentioned than negative impacts; the literature informing the case vignettes did not identify any negative impacts, and most interviewees did not think there were any negative impacts at all. When discussing negative impacts interviewees referred to *potential* negative impacts, but not to actual examples. For instance, three interviewees thought that negative impacts could occur if pharmaceutical companies breached personal data laws or if they used the health data for other purposes than research (INT4, INT8, INT15). If such breaches of law and trust happened, this could negatively impact on public and other stakeholders' trust. Members of the public and potential collaborators could think that the pharmaceutical industry is only using the health data for commercial reasons, which could in turn lead to more pressure on industry or regulations determining that industry has to pay for health data (INT8, INT9). Moreover, if hospital partners do not trust pharmaceutical companies, they may not share health data (INT12). Two interviewees also thought that pharmaceutical companies working with health data could conduct poor-quality analyses, so for example 'fish' for the results they want to find (INT4, INT5), which could 'undermine the credibility of real-world data and its potential benefits for patients' (INT4). This concern was also reported in the literature: companies may conduct analyses using different approaches until they have found a desired result; repeating analyses would be easier than in clinical trials, which are restricted by strict and authorised protocols (Klonoff 2019). An interviewee also felt that if poor-quality health data were used, this could

lead to poor research results and variable evidence, and that it is important to always keep this in mind when working with real-world data (INT14).

4. What are the enablers and barriers to reusing health data?

This chapter articulates the main factors that enable and constrain the effective reuse of health data by the pharmaceutical industry. As with the previous chapter, the analysis presented here is informed by the following sources of evidence: the literature review, the case vignettes (referenced using unique identifiers CV1, CV2, etc.) and stakeholder interviewee insights (as before, the evidence from the interviews has been anonymised and is cited in this chapter using unique interview identifiers (INT1, INT2, etc.)). The enablers and barriers have been identified in this chapter in terms of what they imply for a set of high-level themes drawn out from the evidence base.

4.1. Characteristics related to the access, quality, accuracy and standardisation of health data can be enablers as well as barriers to its reuse

Key takeaways

Reuse of health data by the pharmaceutical industry is hindered by:

- Restrictions on data, meaning that it is not accessible, or only accessible to academic research teams.
- Poor-quality data, i.e. data containing missing data and inaccuracies.
- Available data not containing all the variables required for analysis, or not being updated as regularly as might be desired (this can occur because the primary purpose of collecting data was not for analysis).
- Unintentionally biased data due to the primary purpose of collecting the data not being for its analysis.
- Lack of standardisation and interoperability across datasets meaning analyses using multiple datasets is difficult.

Reuse of health data by the pharmaceutical industry is enabled by:

- Data that are easily accessible to the pharmaceutical industry, e.g. by means of governance procedures that support access.
- Data that are relatively inexpensive to access (this can influence which data are used, e.g. EHR data can be cheaper than prescription and claims data).
- Partnerships with data holders or other bodies who are able to access data.
- The availability of high-quality, curated datasets.
- Secondary analysis tending to cost less than traditional analyses, e.g. secondary analysis can be cheaper than conducting a randomised clinical trial.

4.1.1. Access to health data is a key enabler for its reuse; conversely, the inability to access data can impede reuse by the pharmaceutical industry and result in delays

Having access to data that contains the required and relevant information for a study the data is intended to be used for is important (Galson and Simon 2016; The Academy of Medical Sciences 2016; CV2, CV3, CV3, CV5, CV7, CV9, CV11, INT15). Easy access to data was reported as one of the main enablers of reusing health data, particularly if it is freely or commercially available and good governance procedures are in place to support access (CV2, CV9, INT4, INT14, INT15), rather than needing to go through a paywall or facing governance processes that are difficult to navigate. For example, the use of the Schizophrenia Outpatient Health Outcome dataset to explore methods of identifying schizophrenia drug monitors was supported by the data being made available to the research team without the need for ethical approval (CV9). Involvement of the data owner in the secondary analysis project is another key enabler identified through the case vignettes. For example, access to the THIN database to explore adverse drug side effects was made possible as the company that owned the data allowed the research team access to the dataset (CV3). In addition, the use of EHR and registry data to explore the feasibility of replacing control arms in RCTs with results from secondary data analysis was made easier as the pharmaceutical company involved conducted the RCTs whose data was used in this study (CV5). One interviewee commented on how the US is 'ahead' in health data analysis compared to Europe, in part because accessing data in the US is relatively 'quicker', allowing for faster analysis (INT5). This may be partly due to the FDA in the US having clear regulations for secondary analysis of health data and are supportive of its use compared to European regulators that do not yet have clear regulations (Khosla et al. 2018, INT4, INT12, INT16).

Conversely, inability to access required data is a key barrier for the pharmaceutical industry to reuse health data and can cause delays to studies (Bate, Reynolds and Caubel 2018; Coorevits et al. 2013; CV4, INT2, INT15, INT16). Accessing data can be a notable challenge for the pharmaceutical industry compared to other sectors as data holders are sometimes resistant to sharing data with this sector – for example, when data holders do not want their data to be used for commercial purposes (INT8, INT12) – or pharmaceutical companies themselves may be resistant to sharing information with other companies (Pushpakom et al. 2019). For example, non-randomised study data were used to assess the effects of schizophrenia treatments but accessing the data proved difficult as a pharmaceutical company had originally agreed to provide the data but pulled out of the project while putting together the proposal to apply for the project (CV4). Legislation can also limit access to data (Pushpakom et al. 2019; Wise et al. 2018). For example, it may not be possible to access health data in a standardised way across EU Member States, meaning data from some countries are inaccessible (European Commission 2016).

4.1.2. The quality of health data available in the ecosystem is variable; this can often pose a significant barrier for the pharmaceutical industry to conducting secondary analyses

The nature of the data can influence the ease with which they can be used for secondary analysis by the pharmaceutical industry. There can be many aspects of health data that make it less than optimal for secondary analysis, which has been reported as one of the main barriers in the literature and across the

interviews and case vignettes (Bate, Reynolds, and Caubel 2018; Camm and Fox 2018; Coorevits et al. 2013; European Commission 2016; European Medicines Agency 2016; US FDA 2018; Kalra et al. 2017; Kelly n.d.; Khosla et al. 2018; Love et al. 2016; Nair, Hsu and Celi 2016; Pushpakom et al. 2019; Miani et al. 2014; Safran 2014; Wang 2011; Yildirim et al. 2016; CV2, CV7, INT2, INT3, INT5, INT9, INT12, INT14, INT15). Low-quality data can include such things as inaccuracies, lack of standardisation/interoperability across datasets, lack of applicability of the data and difficulties in understanding the data, each of which are discussed in more detail in below.

Firstly, the quality of the data can impact whether it can be used for secondary analysis. Low-quality data are frequently reported as a barrier with two interviewees suggesting it is a significant barrier faced when reusing health data (INT2, INT12). However, if health data are of high quality, this can act as a real enabler of secondary analysis and makes it more attractive for the pharmaceutical industry to get involved (Makady et al. 2017; INT5, CV7). For example, the Fabry Outcome Survey register was used in a study (CV7) to explore the impacts of Fabry disease on patients due to it being of high quality as it is automatically checked when data are uploaded and clarifications can be sent to the data managers to fill in any gaps. With respect to the quality of data, one case vignette (CV10) related this to EHR data and the aspects of this type of data that supported its use. This case vignette used EHR data as it can be better generalised to the patient population compared to studies of small population sizes which are most often used for schizophrenia research as it provides data on the whole schizophrenia population in the South London and Maudsley area (CV10).

4.1.3. Inaccurate health data can make it difficult for the pharmaceutical industry to extract, interpret and analyse

Inaccuracy is a key issue contributing to poor-quality data. Data that are important for analysis can be missing, such as wider socioeconomic data (CV2, CV7, INT12), or data can be recorded inaccurately, which means certain patients can be included or excluded from studies who should not be (CV8, INT12). Similarly, although data may be recorded, other important patient demographics, such as diet, BMI or ethnicity, may not be which reduces the accuracy of the analysis as assumptions have to be placed on the data (CV2, CV8). For example, when using CPRD to investigate the treatments prescribed to different diabetic populations, the analysis was made difficult as important data could have been missing and other useful patient data, such as ethnicity and diet, is not always collected which means assumptions have to be placed on the data (CV2). In addition, exploration of the link between BMI and risk of NAFLD was undertaken using the THIN and Humedica EHR datasets, however, relevant patient data were missing, such as BMI and smoking status, which meant these patients had to be excluded from the study. Patient alcohol intake may have been incorrectly recorded in the dataset leading to certain patients being incorrectly included or excluded in the study (CV8).

These problems link to an overarching issue in that data used for secondary analysis were originally not collected for that purpose and so it can be difficult to apply in other situations, which can lead to challenges with analysis (European Commission 2016; Galson and Simon 2016; Singh et al. 2018; INT2, INT13, CV11). For example, the use of eight health data sources to test a data derivation tool faced barriers as the data sources used were originally collected for a different purpose and were difficult to apply to the context of the research and required additional assumptions to be placed on the data (CV11).

Even if there are no gaps in the datasets, the literature identifies that datasets might not be updated on a regular basis (Bate, Reynolds and Caubel 2018; Coorevits et al. 2013), meaning the information is outdated and again, may be inaccurate and not reflect the 'real world' situation.

4.1.4. Unintended biases in health data can have a negative impact on secondary analysis

Other aspects of health data can mean it is biased, reducing the accuracy of any analysis conducted on it. For example, patients with more severe symptoms are more likely to be enrolled in the Fabry Outcome Survey registry, which could have led to biased analysis when using the dataset to explore the impacts of Fabry disease on patients. This dataset also has missing data and uncertain data quality as standard assessments of the data quality were not conducted during the study (CV7). In addition, dataset populations are often smaller than real-world populations, which makes it difficult to generalise the data and there may not be long-term follow-up of patients (CV3, CV8). For instance, only one EHR database, THIN, was used in a study to identify adverse drug side effects, making it difficult to generalise across other EHR sources and to other country populations as it only includes data from UK patients (CV3).

Related to this, rare diseases are often poorly understood and described in the real-world population which can make analysis difficult (CV6). This was seen during a study using disease registries to understand the disease progression and treatment of paediatric PAH faced challenges as it is a rare disease which has not been well described and understood in the affected population (CV6).

In addition, health data can have unrecognised variables (European Commission 2016; Klonoff 2019; Singh et al. 2018; INT2, INT13, CV11). Such unrecognised variables include treatment selection bias resulting from, for example, patients requesting a specific treatment or doctors preferring one treatment over the other, or patient characteristics not captured in the data, which may have an impact on a treatment's effect or a disease's progression (e.g. smoking, alcohol use, non-prescribed drugs, lifestyle) (Klonoff 2019).

4.1.5. The current lack of standardisation and interoperability across datasets creates constraints for the secondary analysis of health data by industry

A lack of standardisation and interoperability across datasets was a frequently mentioned barrier, particularly in the literature. Lack of standardisation can contribute to the issue of data inaccuracy discussed previously and a difficulty in cross-analysing different datasets (Coorevits et al. 2013; European Commission 2016; European Medicines Agency 2016; US FDA 2018; Kalra et al. 2017; Kelly n.d.; Khosla et al. 2018; Love et al. 2016; Nair, Hsu, and Celi 2016; Pushpakom et al. 2019; Miani et al. 2014; Safran 2014; Wang 2011; Yildirim et al. 2016; INT2; INT3, INT8, INT9, INT14, INT15) For example, the literature specifically mentions that EHR data sources are often not collected or stored in a consistent manner (Coorevits et al. 2013; Kalra et al. 2017; Nair, Hsu and Celi 2016). Although interoperability across datasets is important, the literature highlighted the complexity of setting this up, adding an additional barrier to reusing health data (for example, different systems are used both within and across countries) (Love et al. 2016; Miani et al. 2014; Singh et al. 2018; INT8, INT9, INT15). However, the European Commission is making progress in this area by focusing on linking health data across EU countries (The London School of Economics and Political Science 2018).

Even if health data are of high quality and accurate, there can be difficulties associated with understanding the data. It is important to present data in an accessible and easily understandable way (Yildirim et al. 2016), however, data used at the moment can be difficult to extract if, for example, it is embedded in handwritten doctors notes (INT2, INT8). Related to a lack of standardisation across datasets, different sets of data use different languages depending on the country the data originates from and terminology which can make it difficult to understand (Marjanovic et al. 2017; The London School of Economics and Political Science 2018). This also contributes to the complexity in creating linked datasets (INT8, INT9, INT15). An interviewee noted, for example, that 'internationally, it's a mess. It's Wild West, there are multiple systems. There are multiple languages, not just human, but machine, vocabulary or knowledge and so on' (INT8).

4.1.6. The secondary analyses of health data tends to be more cost-effective than primary studies, which is perceived as a key enabler to reuse

The costs involved can influence the use of real-world data and what types of data are used. Data that are inexpensive to access can support the reuse of health data, although one interviewee noted that speed of accessing the data may be more important to pharmaceutical companies compared to cost (INT5). Evidence suggests that EHR data are less expensive to access compared to other data types, such as prescription and claims data (Love et al. 2016). Despite the cost that can be involved in purchasing health data, secondary analysis is often much cheaper than conducting a randomised clinical trial. One interviewee suggested that pharmaceutical companies could save money by analysing real-world data compared to using RCTs (INT10).

4.2. Administrative factors need to be carefully considered in the secondary analyses of health data; effective collaboration between the pharmaceutical industry and other sectors is a key enabler of reuse

Key takeaways

Reuse of health data by the pharmaceutical industry is hindered by:

- Administrative factors associated with project structure and management, e.g. staff turnover and slow decision making, particularly on collaborative projects across multiple locations.
- Upfront start-up costs to acquire the skills and infrastructure needed to effectively reuse health data (these costs can disincentivise senior management from investing in reuse of health data).

Reuse of health data by the pharmaceutical industry is enabled by:

- Effective collaborations with a range of stakeholders, both to provide access to data and to provide different ideas, perspectives and skills.
- Cultures within pharmaceutical firms that are open to and accepting of the value of the reuse of health data.

4.2.1. Factors such as team size, staff turnover and lack of funding can be barriers to reusing health data

The case vignettes demonstrated some of the administrative barriers that can make the reuse of health data by the pharmaceutical industry more difficult. For example, a small team size and frequent staff turnover due to the long project length slowed the progress of a study exploring the side effects of drugs using the THIN database (CV3). Funding for this study was also limited which meant the project team had to conduct a lot of the work remotely, rather than face-to-face, which added difficulty in progressing with the project (CV3). Similarly, a large consortium was involved in a project to test a data derivation tool using eight different data sources. Although this was beneficial in providing a range of skills and perspectives to the project, it made management and decision making difficult and slowed progress (CV11).

Initial upfront costs to conduct secondary analysis can act as a barrier, for example, hiring staff with the relevant expertise and investment in the required technology (Kalra et al. 2017; Nair, Hsu and Celi 2016; Wise et al. 2019). It may take a long time to get a return on this initial investment which may act as a disincentive for senior management within pharmaceutical companies from investing (Kalra et al. 2017; Nair, Hsu, and Celi 2016). It was also noted in the literature that there is a lack of sustainable funding for the use of real-world data (European Commission 2016; Nair, Hsu and Celi 2016; Pushpakom et al. 2019).

4.2.2. Cross-organisational and cross-sectoral collaborations are important factors that enable the access to and secondary analyses of health data

Collaboration was seen as key in secondary analyses of health data, as they can provide access to different datasets needed. However, collaboration can cause some difficulties. If there are too many individuals or groups involved – which can occur when several different datasets are needed and therefore several individuals or groups are part of a study – progress can slow down (CV11). However, in general, interviewees noted the importance of collaboration in providing a range of different ideas, perspectives and skills, and identifying strengths and weaknesses in the data, allowing better insights to be gained for the available data and a higher quality output (CV3, CV11, INT5, INT8, INT15). Effective collaboration can support a project in the early stages by ensuring a detailed project plan is produced, standard approaches are taken to analyse the data and expectations are managed across groups early on (CV6). If effective collaborations lead to creation of data infrastructures across organisations, this can enable the processes needed to access the data (INT5, INT8, INT15).

Effective collaboration is particularly beneficial in terms of accessing the data and conducting analysis. This was notably evident through data gathered from the case vignettes, with four of these identifying good collaboration between groups to be a key factor in the project success (CV1, CV3, CV6, CV11). The creation of VISTA, a dataset containing archived clinical trial data on strokes, was supported through collaboration between multiple pharmaceutical companies and a university partner (CV1). In addition, in a study using prospective and observational disease registries to identify a more effective way of evaluating the disease progression and treatment of paediatric PAH, effective collaboration between the pharmaceutical company and the EMA ensured a detailed analysis plan was developed and a standard approach to analysing each of the four datasets to be used was made clear (CV6). Finally, a study testing a

data derivation tool to identify individuals with type II diabetes benefitted from a collaboration of many partners who contributed diverse and useful ideas to the project, helping to identify the strengths and weaknesses of the four sources of data used (CV11).

Collaboration is also beneficial at the later stages of a project during analysis of the data. One case vignette highlighted the benefit of including an individual familiar with the healthcare system the data originates from during analysis of the data (CV3). The study team for this project were investigating the use of the THIN database to identify adverse drug reactions included a senior medical doctor from the UK; by doing this, the team had a greater ability to understand and interpret the THIN database which is a UK primary care dataset (CV3).

In addition to effective collaboration across organisations, having motivated and driven individuals open to the idea of reusing health data involved in a project was deemed as important (CV2, CV12, INT1, INT2). For example, a pharmaceutical company and a university involved in understanding the populations prescribed glucose-lowering drugs using the CPRD database were motivated and interested in the project which drove it forward (CV2). Within pharmaceutical companies, it was noted that having a culture open to and accepting of real-world data supported the company in reusing health data (INT2). To illustrate this point, when using EHR data to obtain NICE approval for reimbursement for a lung cancer treatment, the pharmaceutical company involved placed value on evidence arising from secondary data analysis and were open to using it for strategic decision making (CV12). This is particularly relevant for those in senior leadership roles as if they recognise the value of reusing health data, they are more likely to provide the required resources (INT1, INT2). Demonstrating that investing in secondary analysis of health data is worthwhile in the long term can help support this (Wise et al. 2019). Having a lack of buy-in or willingness to take risks on secondary analysis is seen as a barrier (Khosla et al. 2018; INT1).

4.3. Lack of uniform regulations and clear guidelines (including issues related to data protection and data privacy) can hinder secondary analyses of health data

Key take-aways

Reuse of health data by the pharmaceutical industry is hindered by:

- A lack of clear and uniform regulations on and information about what is and is not acceptable in terms of health data reuse. (This includes whether secondary analysis can be submitted as valid evidence for a drug's efficacy, how health data are allowed to be used by the pharmaceutical industry, and what is legal and acceptable in terms of secondary analysis.)
- The existence of intra- and inter-country differences in approaches to regulations and guidelines for secondary analysis.
- HTA bodies not recognising evidence generated through secondary analysis of health data.
- Lack of clarity on ownership of outputs of secondary data analysis, and who has the right to share data and with whom.
- Lack of clarity in relation to GDPR and the systems that need to be in place to ensure effective data governance and protection of health data.
- Only being able to access anonymised health data which may have had desired variables removed.

Reuse of health data by the pharmaceutical industry is enabled by:

- Regulators and policymakers recognising the value of and demanding real-world evidence before they make decisions.

Some literature reports that regulators and policymakers are increasingly demanding real-world evidence and seeing the value in it which acts as a motivator for pharmaceutical companies to get involved in secondary analysis of health data (Khosla et al. 2018). However, across the literature and interviews, it is clear that a lack of clear and uniform regulations is a barrier to the pharmaceutical industry engaging in health data reuse. Interviewees frequently commented that they felt that current guidance and regulation from regulators are not clear on firstly, whether secondary analysis of data can be submitted as valid evidence for a drug's efficacy and secondly, how health data can be used by the pharmaceutical industry and what rules should be followed in conducting secondary analysis (The Academy of Medical Sciences 2016; INT4, INT12, INT16). This lack of clarity is made worse by the diverse views and approaches of different country- and European-level regulators in providing guidance on secondary analysis. There is no top down mandate and it is unclear which of the guidelines from different regulators should be followed (Khosla et al. 2018; INT16). For example, an interviewee felt that 'having scientific advice, for example from the European Medicines Agency, on questions related to the evidentiary value of real-world studies conducted using real-world data would be important' (INT4). Similarly, another interviewee highlighted that 'a clear signal [from regulators] regarding acceptance would be good in the future' (INT2; insert added). In addition, many European countries' HTA guidelines do not recognise evidence generated through secondary analysis of health data as robust enough to be submitted in HTA processes, but also may not be accepted due to lack of patient involvement (The London School of Economics and Political Science 2018).

Analytical abilities and potential uses of health data are expanding and developing faster than regulations are being published, which may be contributing to this challenge, although it was noted in interviews that

this situation is improving (INT14, INT16). It may also be because regulators and policymakers do not fully understand real-world evidence and how it can be used by the pharmaceutical industry, for example having uncertainty over the reliability of secondary analysis due to lack of randomisation (Khosla et al. 2018), thus they may not place value on it as a robust source of evidence.

The lack of a uniform regulation or guidance for conducting secondary analysis of data may, in part, be due to the difficulty of implementing a single approach across all EU Member States. There is large variation in how data are handled across EU Member States, with different initiatives in place to support reuse of data and interoperability and different health systems (Coorevits et al. 2013; European Commission 2007, 2016; European Medicines Agency 2016; Miani et al. 2014; Yildirim et al. 2016; INT10, INT15). For example, one interviewee provided the example that the UK has a more open (access) approach to reusing health data, such as that seen with the THIN database, whereas Germany is more restrictive in who it allows to access data (INT10). However, even within the UK (for example) there are differences in general access to health data: for example, Scottish and Welsh health organisations tend to be more restrictive in terms of sharing health data than health organisations in England (Lyons 2014; McCartney 2014). Moreover, Scotland and Wales have better privacy and data protection policies in place in that they deidentify data, while the UK Health and Care Social Care Act 2012 stipulates that general practices in England are required to upload raw data which may be shared with researchers (McCartney 2014).

Legal barriers can also occur when reusing health data (INT12, INT15), for example, deciding who owns the intellectual property of the output of the secondary data analysis, who owns the data, and who has the right to share it and with whom (Pushpakom et al. 2019; INT12). The European Commission reported that legal and operational barriers could be important when using registry data (European Commission 2016). An interviewee also mentioned that as there are different governance structures, rules and laws in different countries, it can be a challenge for companies working in various countries to conduct secondary analyses in an efficient way (INT15). The issue of heterogeneity within and between countries was also highlighted in the literature (Cole and Towse 2018).

One interviewee raised the challenge of governance structures put in place by the data owners being the same for all study types, regardless of whether they are low or high risk. This can make access to the data more difficult, which can slow down the progress of low-risk, descriptive studies that should not require such strict access rules (INT16).

Ensuring effective data protection and confidentiality processes are in place can be difficult (Pushpakom et al. 2019; Wise et al. 2018). For example, the consent required and the validity of informed consent from patients can be unclear (Cole and Towse 2018; Safran et al. 2007; The Academy of Medical Sciences 2016; Yildirim et al. 2016; INT2) and scaling up datasets to allow the data to be shared with a larger number of organisations while maintaining rigorous data protection approaches can be challenging (Kelly n.d.). It is important that the data pharmaceutical companies access are anonymised and/or de-identified, particularly with datasets of small patient populations as these can easily be traced back to the individual (The London School of Economics and Political Science 2018; INT2). However, although anonymisation of data is important, this can lead to difficulties with analysing it as useful data can be

removed (e.g. one interviewee mentioned the point of removing patient location data which can be used to assess their access to healthcare (INT2)).

Interviewees and literature sources highlighted the lack of clarity with data protection after GDPR was introduced in 2018. It has led to some open questions about the processes that need to be in place to ensure effective data governance and protection, making it difficult for pharmaceutical companies to decide on the best practice for data protection and may mean they are fearful of using the data, seeing it as too risky due to the possible legal repercussions it could result in (The London School of Economics and Political Science 2018; INT3, INT4).

4.4. Analytical skills and capabilities within the pharmaceutical industry are crucial to enabling effective secondary analyses of health data

Key takeaways

Reuse of health data by the pharmaceutical industry is hindered by:

- Lack of development of methods for reusing health data.
- Lack of skills and experience at pharmaceutical companies to effectively analyse health data.

Reuse of health data by the pharmaceutical industry is enabled by:

- Continual development of new tools and methods to reuse health data.
- Pharmaceutical companies having access to staff – either in-house or externally – with adequate skills to conduct high-quality analyses, interpret the data correctly and/or judge the accuracy of health data.

Once pharmaceutical companies have access to the relevant data, they need to have the required analytical skills, either in-house or externally, to conduct high-quality analysis, interpret the data correctly and judge the accuracy of health data (INT2, INT4, INT5, INT15). Three case vignettes highlighted the importance of this, as well as having the skills to overcome any challenges presented by the data (CV2, CV3, CV12). Firstly, using CPRD to explore the treatments prescribed to different diabetic populations was supported by a pharmaceutical company having previous experience in using and analysing this dataset (CV2). Secondly, the broad consortium of organisations involved in a study using the THIN database to identify adverse drug reactions meant they had the required skillset to conduct the research to a high quality (CV3). Finally, the pharmaceutical company that was involved in a project using EHR data to obtain NICE approval for reimbursement for a lung cancer treatment had the staff knowledge and ability to conduct secondary analysis and overcome any problems that may have arisen with the data (CV12).

Limitations in analytical tools and interpreting data can be barriers to effective analysis which may be due to real-world big data analysis still being a relatively new approach (Miani et al. 2014; Singh et al. 2018). It may result in some parts of the data not being analysed at all or subjective analysis takes place, for example, if the tool used has not been verified (CV3, CV11). For instance, limitations in an exploratory analysis of the THIN database to identify adverse drug events meant it was difficult to evaluate false negatives (CV3). Additionally, the algorithm used to identify patients with type II diabetes in a study testing a data derivation tool using a variety of health data sources had not been validated and relied on the judgement of experts during its use which may have been subjective (CV11). There is also no standard

approach taken to analyse health data (Miani et al. 2014). Limited tools to analyse the data can subsequently lead to difficulties in interpreting the data, such as identifying causality between a drug and an effect, and the potential for the loss of important contextual information when data are aggregated (Bate, Reynolds and Caubel 2018; Biotechnology Industry Organization 2011; Nair, Hsu and Celi 2016; Safran 2014).

Pharmaceutical companies may lack the skills to effectively analyse health data (Pushpakom et al. 2019; The Academy of Medical Sciences 2016; INT1, INT14). There can also be varied approaches to assessing the quality of the data before analysis, which can mean the quality of results varies (CV7). For example, when using the Fabry Outcome Survey registry to explore the impact of Fabry disease on female patients, there was a lack of standardised assessment of the data across the different groups involved in the analysis which may have had implications for the analysis (CV7). In addition to a lack of analytical skills, pharmaceutical companies may not have the staff with the appropriate knowledge of real-world data. This can lead to misunderstandings about when EHR data, for example, could be used (INT9). To overcome this, it is important to communicate the studies where RCTs are more difficult to conduct, such as with children or on co-morbidities, and how secondary analysis can be used as a complementary source of evidence in these situations (The London School of Economics and Political Science 2018; The Academy of Medical Sciences 2018). There can also be misjudgements in the best data to use for a particular task, resulting in inappropriate data being used, for example, industry may incorrectly assume that data from the US are 'better' than that available in Europe (INT5). This may be because pharmaceutical companies are relying on data from the US that they have used before and so are familiar with it and know it is high quality, which may lead to these companies being cautious about using a different dataset from Europe they have not used before (INT5).

Although it is important for pharmaceutical companies to have the skills to analyse data, one interviewee discussed the need for data suppliers to have the skills and knowledge to interpret their own data, understand how it is collected and where the gaps might be (INT5).

Technological advances and creation of new technology has made it easier to capture, anonymise, standardise and analyse a greater amount of data (Love et al. 2016; Safran et al. 2007). For example, new tools have been developed that can accelerate data analysis, such as a health data tool developed by a health analytics company that enables pharmaceutical companies to run rapid cycle analytics to generate insights from health data quicker than other analytical methods, which can then be used for strategic decision making (INT16). Other tools support extraction of data from unstructured sources, for example, from handwritten doctors notes, to create a structured dataset (Love et al. 2016).

4.5. Public and healthcare provider trust in pharmaceutical companies accessing and using health data currently poses a challenge for the pharmaceutical industry

Key takeaways

Reuse of health data by the pharmaceutical industry is hindered by:

- Lack of public and healthcare provider trust due to concerns over private companies accessing personal health data and not having adequate data protection processes in place.
- Concerns in pharmaceutical companies over lack of public trust and potential loss of public trust.

Reuse of health data by the pharmaceutical industry is enabled by:

- Clear governance processes detailing how data can be used and who it can be used by.
- Only using anonymised datasets in the secondary analyses of health data.

The interviews and literature highlight that public trust is a key factor in supporting pharmaceutical companies accessing data. However, there is often a lack of public trust due to concerns over private companies accessing personal health data and not having adequate data protection processes in place (Bate, Reynolds and Caubel 2018; Heath 2010; Nair, Hsu and Celi 2016; Pushpakom et al. 2019; Miani et al. 2014; The Academy of Medical Sciences 2016; Yildirim et al. 2016; INT3, INT4, INT5, INT8, INT15). Some interviewees thought the lack of public trust is due to inadequate communication with the public about how pharmaceutical companies use health data and some of the benefits that can results from reuse of data (INT4, INT8, INT14, INT15). It is often the case that the public is only made aware of 'mistakes', such as data breaches or use of data for an improper reason, rather than the positive impacts it can have (INT5, INT8). An interviewee felt that 'there only needs to be one mistake of data loss or revealing or misuse, etc. for companies to no longer be able to access the data' (INT5). Another interviewee explained that there is often misunderstanding of what the pharmaceutical industry does, noting that there is a 'challenge that people have in understanding the nuances of pharma or industry organisations. If you talk about pharma, they think it is commercial, sales, marketing. That's one part. But what they don't take into account is the rest of the organisation, that we do R&D, etc.' (INT8).

One interviewee argued that anonymising data can reassure the public that pharmaceutical companies are not going to be able to identify them from the dataset (INT5). Anonymising data also enables it to be shared across organisations (INT1). Clear governance structures are important in safeguarding data (INT4). For example, having clear boundaries in place for what pharmaceutical companies can and cannot do with the data, such as is seen with the VISTA database (CV10). Pharmaceutical companies are able to input data into the centralised VISTA database (an archive of stroke trial data) as they are not allowed to reanalyse the data for market gain, e.g. patent applications or regulatory submission (CV1).

Other organisations may not wish to share data with pharmaceutical companies due to similar concerns to those held by the public (INT8, INT15). This issue is particularly prevalent within healthcare systems as healthcare professionals may be resistant to sharing their patients' data with private companies due to concerns over misuse (Miani et al. 2014; INT16). One interviewee noted that, in the same way that there is no single, clear regulation for pharmaceutical companies to follow when reusing health data, there is

also not a united approach to improving the trust healthcare professionals have in pharmaceutical companies (INT16).

In addition, data partners and data owners also need to have trust in pharmaceutical companies that they will use the data correctly (INT12).

5. Reflections on the future of the reuse of health data

In the report so far, we have articulated how the pharmaceutical sector reuses a variety of health data. In addition, we have drawn out the key enablers and barriers to the reuse of health data. Using this information, we offer some reflections in this chapter on the future direction of the reuse of health data by the pharmaceutical sector. In this regard, the literature and the stakeholders we interviewed in this study have suggested a number of topics for further consideration by the pharmaceutical sector as well as wider stakeholders involved in the health data ecosystem. Some of the issues discussed in this chapter have clear implications for strategies and actions in the future. We have proposed a set of wide-ranging ideas or topics which need to be considered by the pharmaceutical industry and other stakeholders to get the most out of the reuse of health data, and in doing so, help create an effective and enabling health data ecosystem. We have specifically considered perceptions of what might happen with the secondary use of health data by pharmaceutical companies in the near- to medium-term future – in the context of this study, this time period refers to future developments that are likely to take place within one to five years. (As has been done in previous chapters, the evidence from the interviews has been anonymised using unique interview identifiers (INT1, INT2, etc.).)

5.1.1. What might the future look like?

Key takeaways

- It is likely that secondary analysis of health data will increase in the near- to medium-term future, expanding to cover different disease areas, new analytical methods and different types of data.
- As they become more involved in the health data ecosystem, patient organisations are likely to play a bigger role in secondary data analysis in the future, which may help overcome some data protection concerns held by the public.
- Regulations and guidelines associated with health data reuse are likely to become clearer and more coordinated in the future, as has been seen recently in reports produced by the US FDA and the EMA.

Interviewees expect secondary analysis of health data to increase in the future, in particular expanding to be used in other areas, for example, outside of major diseases, such as cancer, and for more rare conditions (INT2, INT8, INT9, INT14). An interviewee felt that already 'in the recent years [secondary use] has become mainstream. I believe that will continue. In one or two years, maybe three-quarters of non-interventional studies will have secondary data.' (INT14). Another interviewee noted that greater reuse of data could drive competition among pharmaceutical companies and lead to improvements in and 'higher quality' extraction of data and analysis (INT2).

Interviewees highlighted the likelihood that currently available data types will be used in different ways in the future and new data types will be used for analysis. Possible uses included real-world data being used more for indirect treatment comparisons to allow existing and new drugs to be compared (INT5), and to target and identify subpopulations in which treatments are more beneficial or less safe (INT7). It was also suggested that secondary analysis may be used more to replace RCTs or alongside RCTs in drug development and approval (INT8, INT14). However, one interviewee felt it was unlikely that secondary analysis will fully replace RCTs; rather it might be used in situations where RCTs are difficult to conduct (INT12). A wider range of data sources could also be used for certain tasks, for example, EHR data are already used for pre- and post-authorisation studies, but it was suggested that registry data could also be used for this purpose in the near future (INT9).

Some interviewees also expect increased use of newer data types such as social media data, the use of which are already being tested. For example, there is ongoing work testing the feasibility of using social media to study the spread of infection (INT5). Similarly, analysis of data collected from wearable devices and health apps was expected to become more commonplace, although this is a less developed area of research for the pharmaceutical industry (European Medicines Agency 2016; INT5). This may result in patients themselves becoming the largest 'generators of data' (INT3, INT15). An interviewee felt that in the near future 'data will be more driven by individuals (i.e. patients) wanting to share' (INT15), as opposed to being aggregated by the health system (i.e. 'less by doctors, hospitals, etc.' (INT15)).

In addition to use in other disease areas, interviewees expect greater involvement of healthcare systems and patient organisations in secondary analysis of health data. One interviewee noted that as healthcare systems collect large amounts of data, there is a need for them to be more involved in secondary analysis and to ensure the data they collect is used efficiently (INT16). Pharmaceutical companies will also be expected to start working with patient organisations more regularly in order to take a more holistic approach to data analysis, obtain greater access to patient reported data, and to give patients a larger role in sharing their own data (INT5, INT9, INT12). It was suggested that this might help overcome some of the issues of public trust and data protection discussed previously (see Section 4.5) as patient organisations could be the 'data owners' and decide how pharmaceutical companies can reuse the data (INT5).

Some interviewees expect that increased reuse of health data will be enabled by development of new technologies and analytical methods, as well as improved interoperability and standards of data (INT10, INT14), and that this, in turn, could also lead to greater acceptance of health data reuse by payers, insurers and regulators (INT14). In addition, as more datasets are merged together, the cost of storing and analysing the data would be expected to drop (Nair, Hsu and Celi 2016).

Regulations and guidelines are also expected to become clearer and more coordinated in the future. The US FDA has already accepted the use of claims data in safety assessments and is in the process of creating clearer guidelines for the use of real-world data in making regulatory decisions in the US (INT12) (The Academy of Medical Sciences 2018). It is also exploring ways to assess the quality of real-world data and methods used to analyse it. In addition, although there have been some difficulties in confidently interpreting GDPR legislation, it could support data protection decisions in the future as all companies across the EU will be following the same law (INT10). Finally, Wise et al. (2019) highlight the

importance of 'FIND' (findable, accessible, interoperable and reusable) data principles for improving data governance and increasing access to data for companies.

5.1.2. What are the priority topics for further discussion that might help create a sustainable ecosystem in which health data is reused effectively?

Key takeaways

- It is important that the pharmaceutical industry actively continues to explore the reuse of different types of health data beyond the data types (such as EHRs, registry data and clinical trial data) that have been traditionally used in secondary analysis.
- Health data reuse is a growing field, both within the pharmaceutical industry and beyond, and there is a need for continued research and development on analytical tools and techniques (including the ability to link different datasets).
- To improve accessibility to health data, a greater degree of collaboration is needed between the pharmaceutical industry and other key stakeholders, such as regulators and healthcare system actors.
- As the health data ecosystem evolves, promoting harmonised standards and interoperability across datasets could enable health data to be used more effectively and efficiently.
- There is a need for clearer and more uniform regulations (including for data protection) and guidelines related to secondary data analysis.
- Improving data and analytical skills within the pharmaceutical industry is key to enabling effective secondary analyses of health data.
- Building public confidence can facilitate buy-in and trust and promote the further reuse of health data by the pharmaceutical industry.

The health data ecosystem is complex, rapidly evolving and consists of several interdependent stakeholders, often with competing priorities. Despite the many technological, regulatory and socioeconomic challenges associated with health data reuse by the pharmaceutical industry, there is general acknowledgement across the different actors of the potential benefits of secondary analysis of health data. In this section, we propose a set of topics for further consideration by the pharmaceutical industry as well as other stakeholders in the health data ecosystem. Inevitably, 'solutions' to some of the issues that have been highlighted in this study would require coordinated and proactive action across multiple stakeholders. Each of these topics is discussed in turn below.

It is important that the pharmaceutical industry actively continues to explore the reuse of different types of health data beyond the data types that have been traditionally used in secondary analysis

The pharmaceutical industry is currently reusing a range of health data, in particular, EHRs, registry data and clinical trial data. As we have seen, the types of health data being used in secondary analysis depend on a number of factors including but not limited to the following: the stage in the innovation pathway; the type of company undertaking the research; the location of the research (e.g. Europe vs US); the methodology; and/or the ability of a company to access the data. Furthermore, as noted in Section 5.1.1, the reuse of health data by the pharmaceutical industry is likely to increase in the short- to medium-term. While EHRs, registry data and clinical trial data (in particular) tend to be the most commonly used data types across the pharmaceutical industry, there are several other less traditional types of data that are also

being used, but to a lesser extent. These include (for example) biobank data prescribing and dispensing data, claims data, social media data and data from wearable devices. It is important that the pharmaceutical industry keeps up to date with the pace of change of developments in these fields with particular regard to building methods and tools to reuse these data effectively and efficiently in order to acquire better insights and potentially capture more value from data.

Health data reuse is a growing field, both within the pharmaceutical industry and beyond, and there is a need for continued research and development on analytical tools and techniques

Across a number of the interviews, it was felt that more research and development is required in relation to acquiring the necessary methodological tools and techniques to support secondary analysis of health data, as well as the ability to link different datasets. This would be particularly relevant in cases requiring improved analysis and/or using larger and more complex datasets. For example, the importance of validating common data models to ensure data sources used for secondary analysis are reliable has been highlighted. In addition, there is a need for the pharmaceutical industry to take advantage of new and emerging technologies that could potentially make (for example) the capture and anonymisation of health data swifter and facilitate a greater amount of data to be analysed. Wise et al. (2019), for instance, discuss the benefits of reading and analysing data with artificial intelligence (AI).

To improve accessibility to health data, a greater degree of collaboration is needed between the pharmaceutical industry and other key stakeholders, such as regulators and healthcare system actors

Evidence from the literature and experts we interviewed strongly supported the need for greater levels of collaboration across the different stakeholders in the health data ecosystem. This included, for example, collaborations with and across regulators, health technology assessors and payers for post-authorisation studies, as well as 'better' and more open relationships between the pharmaceutical industry and healthcare systems to support the sharing of data between both sectors. Collaborations are viewed as a particularly valuable way for the pharmaceutical industry to gain access to a variety of health datasets that typically might not be accessible, as well as to provide different ways to conducting analyses. In addition, multi-stakeholder collaborations serve as a means to obtaining the necessary skills to analyse health data, and also make a range of perspectives and ideas available that might otherwise not be available to the pharmaceutical industry. In general, greater collaboration and coordination across the different stakeholders could facilitate a greater degree of health data reuse.

As the health data ecosystem evolves, promoting harmonised standards and interoperability across datasets could enable health data to be used more effectively and efficiently

Despite being a potentially complex process, the interviews and literature highlighted the importance of promoting the development and adoption of harmonised standards in (for example) data collection and in ensuring seamless interoperability across datasets to facilitate better cross-analysis of datasets. Developing clear, unambiguous and interoperable standards could contribute to improving data accuracy and

facilitate the cross-analyses of different datasets. Standardising datasets could also have other benefits such as reducing the time needed for manual data entry (Kelly n.d.), but may require financial investment. Furthermore, linking different datasets could allow for a more comprehensive view of a patient's healthcare pathway, which could lead to greater demand and usability of data.

There is a need for clearer and more uniform regulations (including for data protection) and guidelines related to secondary data analysis

Across a number of the interviews, it was noted that regulators could develop clearer regulations and guidelines on the reuse of health data (e.g. what is and what is not acceptable real-world data evidence). This would promote clarity as to the value of secondary analyses in different circumstances. Furthermore, developing uniform regulations and guidelines could reduce inter-country differences in approaches to secondary analysis of health data. This is particularly important in the context of data protection laws to ensure data protection processes within pharmaceutical companies comply with legislation. However, current data protection laws, such as GDPR, are sometimes interpreted differently across EU Member States. As noted above, standards are needed to ensure secondary analysis is effective and efficient; for example, there could be a unified set of 'goalposts' that regulators decide on for industry to meet (Wise et al. 2018; INT3, INT16). Finally, it was also felt that regulators could drive the growth of the reuse of health data by developing greater acceptance of real-world data as evidence

Improving data and analytical skills within the pharmaceutical industry is key to enabling effective secondary analyses of health data

As discussed previously (Section 4.4), the lack of data and analytical skills within the pharmaceutical industry was generally acknowledged across several interviews as an important barrier to reusing health data. It was highlighted that there is a need for training and upskilling to build and improve analytical skills to ensure there is a sufficiently large workforce within the pharmaceutical industry to conduct secondary analysis. This includes roles such as data scientists, epidemiologists and health economists. Pharmaceutical companies could benefit by investing in the infrastructure, tools and expertise to ensure they are able to analyse data to a high standard, for example, by training their existing staff or hiring new staff with the relevant experience and implementing better data management systems. As well as the pharmaceutical industry, other actors in the wider health system could benefit from these skills, e.g. clinicians having the ability to use the data they collect for secondary analysis.

Building public confidence can facilitate buy-in and trust and promote the further reuse of health data by the pharmaceutical industry

As noted previously (Section 4.5), there has traditionally been a lack of public (and healthcare provider) trust of the pharmaceutical industry's reuse of health data. To build public confidence, it may help to increase transparency in how health data are used by the pharmaceutical industry and share examples of positive impacts that have resulted from the pharmaceutical industry reusing health data.

References

ABPI. 2007. 'ABPI Guidelines for the Secondary Use of Data for Medical Research Purposes'. London: ABPI. As of 11 August 2019: https://www.onlymedics.com/documents/abpi-guidelines-data-for-medical-research.pdf.

Ali, Myzoon, Rachael MacIsaac, Terence J. Quinn, Philip M. Bath, David L. Veenstra, Yaping Xu, Marian C. Brady, Anita Patel & Kennedy R. Lees. 2017. 'Dependency and Health Utilities in Stroke: Data to Inform Cost-Effectiveness Analyses'. *European Stroke Journal* 2 (1): 70–76. https://doi.org/10.1177/2396987316683780.

Ankarfeldt, Mikkel Z., Brian L. Thorsted, Rolf H.H. Groenwold, Erpur Adalsteinsson, M. Sanni Ali & Olaf H. Klungel. 2017. 'Assessment of Channeling Bias among Initiators of Glucose-Lowering Drugs: A UK Cohort Study'. *Clinical Epidemiology* 9: 19–30. https://doi.org/10.2147/CLEP.S124054.

Bate, Andrew, Robert F. Reynolds & Patrick Caubel. 2018. 'The Hope, Hype and Reality of Big Data for Pharmacovigilance'. *Therapeutic Advances in Drug Safety* 9 (1): 5–11. https://doi.org/10.1177/2042098617736422.

Biotechnology Industry Organization. 2011. 'FDA's Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data Sets'. BIO. As of 11 August 2019: https://www.bio.org/advocacy/letters/fdas-best-practices-conducting-and-reporting-pharmacoepidemiologic-safety-studies-u.

Camm, A. John, & Keith A. A. Fox. 2018. 'Strengths and Weaknesses of "Real-World" Studies Involving Non-Vitamin K Antagonist Oral Anticoagulants'. *Open Heart* 5 (1): e000788. https://doi.org/10.1136/openhrt-2018-000788.

Cederholm, S., G. Hill, A. Asiimwe, A. Bate, F. Bhayat, G. Persson Brobert, T. Bergvall, D. Ansell, K. Star, & G. N. Norén. 2015. 'Structured Assessment for Prospective Identification of Safety Signals in Electronic Medical Records: Evaluation in the Health Improvement Network'. *Drug Safety* 38 (1): 87–100. https://doi.org/10.1007/s40264-014-0251-y.

ClinicalTrials.gov. 2019. 'Glossary of Common Site Terms'. As of 11 August 2019: https://clinicaltrials.gov/ct2/about-studies/glossary.

Cochrane Library. n.d. 'Cochrane Reviews'. As of 11 August 2019. https://www.cochranelibrary.com/.

Cole, Amanda, & Adrian Towse. 2018. 'Legal Barriers to the Better Use of Health Data to Deliver Pharmaceutical Innovation'. London: Office of Health Economics.

Collins, H. 2016. 'Hungary Hopes Health App Will Launch New Era of Care'. politico.eu, 29 November 2016, 12.00 p.m. CET. As of 11 August 2019: https://www.politico.eu/article/hungary-hopes-health-app-will-launch-new-era-of-care/.

Coorevits, Pascal, Mats Sundgren, Gunnar O. Klein, Anne Bahr, Brecht Claerhout, Christel Daniel, Martin Dugas, et al. 2013. 'Electronic Health Records: New Opportunities for Clinical Research'. *Journal of Internal Medicine* 274 (6): 547–60. https://doi.org/10.1111/joim.12119.

CPRD. 2019. 'Data Access'. As of 11 August 2019: https://cprd.com/Data-access.

CPRD (homepage). n.d. As of 11 August 2019: https://www.cprd.com/.

Efthimiou, Orestis, Dimitris Mavridis, Thomas P. A. Debray, Myrto Samara, Mark Belger, George C. M. Siontis, Stefan Leucht & Georgia Salanti. 2017. 'Combining Randomized and Non-Randomized Evidence in Network Meta-Analysis'. *Statistics in Medicine* 36 (8): 1210–26. https://doi.org/10.1002/sim.7223.

EU Clinical Trials Register. n.d. 'Clinical Trials Register'. As of 11 August 2019: https://www.clinicaltrialsregister.eu/ctr-search/search.

European Commission. 2007. 'Primary and Secondary Use of EHR Systems: Enhancing Clinical Research for Better Health and High Quality Healthcare.'

European Commission. 2016. 'STAMP Commission Expert Group. 10 March 2016. Subject. Real World Evidence. Agenda Item 4'. STAMP 4/23 rev.1. As of 11 August 2019: https://ec.europa.eu/health//sites/health/files/files/committee/stamp/2016-03_stamp4/4_real_world_evidence_background_paper.pdf.

European Medicines Agency. 2016. 'Identifying Opportunities for "Big Data" in Medicines Development and Regulatory Science. Report from a Workshop Held by EMA on 14–15 November 2016'. London: European Medicines Agency. As of 11 August 2019: https://www.ema.europa.eu/en/documents/report/report-workshop-identifying-opportunities-big-data-medicines-development-regulatory-science_en.pdf.

European Network of Cancer Registries (homepage). n.d. As of 11 August 2019. https://www.encr.eu/.

European Society for Medical Oncology. 2018. 'Large National Sequencing Projects in Europe'. esmo.org, 25 April 2018. As of 11 August 2019: https://www.esmo.org/Oncology-News/Large-National-Sequencing-Projects-in-Europe.

European Union. 2016. 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)'. As of 11 August 2019: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN.

Flatiron. n.d. 'Confidentiality Commitments'. Flatiron Health. As of 11 August 2019: https://flatiron.com/confidentiality-commitments/.

Franzén, Stefan, Christer Janson, Kjell Larsson, Max Petzold, Urban Olsson, Gunnar Magnusson, Gunilla Telg, Gene Colice, Gunnar Johansson & Mats Sundgren. 2016. 'Evaluation of the Use

of Swedish Integrated Electronic Health Records and Register Health Care Data as Support Clinical Trials in Severe Asthma: The PACEHR Study'. *Respiratory Research* 17 (1): 152. https://doi.org/10.1186/s12931-016-0461-1.

Galson, Steven, & Gregory Simon. 2016. 'Real-World Evidence to Guide the Approval and Use of New Treatments'. *Perspectives: Expert Voices in Health & Health Care*, 18 October 2016. As of 11 August 2019: https://nam.edu/wp-content/uploads/2016/10/Real-World-Evidence-to-Guide-the-Approval-and-Use-of-New-Treatments.pdf.

Giugliani, Roberto, Dau-Ming Niu, Uma Ramaswami, Michael West, Derralynn Hughes, Christoph Kampmann, Guillem Pintos-Morell, Kathleen Nicholls, Jörn-Magnus Schenk & Michael Beck. 2016. 'A 15-Year Perspective of the Fabry Outcome Survey'. *Journal of Inborn Errors of Metabolism and Screening* 4 (January): 2326409816666298. https://doi.org/10.1177/2326409816666298.

Gliklich, Richard E., Nancy A. Dreyer & Michelle B. Leavy. 2014. 'Analysis of Linked Registry Data Sets'. In *Registries for Evaluating Patient Outcomes: A User's Guide*, edited by Richard E. Gliklich, Nancy A. Dreyer & Michelle B. Leavy, 3rd ed. Rockville, MD: Agency for Healthcare Research and Quality.

Health Economics Research Centre. n.d. 'Clinical Practice Research Datalink'. As of 11 August 2019: https://www.herc.ox.ac.uk/downloads/health_datasets/browse-data-sets/general-practice-research-database-gprd.

Heath, Jennifer. 2010. 'Emerging Consumers View of Secondary Uses of Medical Data'. In *2010 IEEE International Symposium on Technology and Society*, 87–95. https://doi.org/10.1109/ISTAS.2010.5514650.

Hernberg-Ståhl, Elizabeth. 2006. 'Organization and Technical Aspects of FOS – the Fabry Outcome Survey'. In *Fabry Disease: Perspectives from 5 Years of FOS*, edited by Atul Mehta, Michael Beck & Gere Sunder-Plassmann. Oxford: Oxford PharmaGenesis. As of 11 August 2019: https://www.ncbi.nlm.nih.gov/books/NBK11596/.

Hesse, Kerrick, Rachael L. MacIsaac, Azmil H. Abdul-Rahim, Patrick D. Lyden, Erich Bluhmki, Kennedy R. Lees & VISTA Collaborators. 2016. 'Online Tool to Improve Stratification of Adverse Events in Stroke Clinical Trials'. *Stroke* 47 (3): 882–85. https://doi.org/10.1161/STROKEAHA.115.011930.

Hughes, Derralynn, Miguel-Ángel Barba Romero, Andrey Gurevich, Patrick Engrand & Roberto Giugliani. 2018. 'Menarche, Menopause, and Pregnancy Data in Untreated Females and Females Treated with Agalsidase Alfa in the Fabry Outcome Survey'. *Molecular Genetics and Metabolism*, Lysosome (2018), 123 (2): S67. https://doi.org/10.1016/j.ymgme.2017.12.164.

Hughes, Jane P., Steve Rees, S. Barret Kalindjian & Karen Philpott. 2011. 'Principles of Early Drug Discovery'. *British Journal of Pharmacology* 162: 1239–49. https://doi.org/10.1111/j.1476-5381.2010.01127.x.

IMI. n.d. 'About IMI'. As of 11 August 2019: http://www.imi.europa.eu/about-imi.

INPDR (homepage). n.d. As of 11 August 2019: https://inpdr.org/.

Kalra, Dipak, Veli Stroetmann, Mats Sundgren, Danielle Dupont, Irene Schlünder, Geert Thienpont, Pascal Coorevits & Georges De Moor. 2017. 'The European Institute for Innovation through Health Data'. *Learning Health Systems* 1 (1): e10008. https://doi.org/10.1002/lrh2.10008.

Kelly, Brian. n.d. 'Registries and the Future of Medicine'. As of 11 August 2019: http://www.pharmexec.com/registries-and-future-medicine.

Khosla, Sajan, Roderick Grenville White, J. Linares Medina, Mario Jnm Ouwens, Cathy Emmas, Tim Koder, Gary Male, Sandra J. Leonard, Mattias Kyhlstedt & Marc L. Berger. 2018. 'Real World Evidence (RWE) – a Disruptive Innovation or the Quiet Evolution of Medical Evidence Generation? [Version 2; Referees: 2 Approved]'. *F1000Research* 111 (7). https://doi.org/10.12688/f1000research.13585.1.

Klonoff, David C. 2019. 'The Expanding Role of Real-World Evidence Trials in Health Care Decision Making'. *Journal of Diabetes Science and Technology*, March, 1932296819832653. https://doi.org/10.1177/1932296819832653.

Loomis, A. Katrina, Shaum Kabadi, David Preiss, Craig Hyde, Vinicius Bonato, Matthew St. Louis, Jigar Desai, et al. 2016. 'Body Mass Index and Risk of Nonalcoholic Fatty Liver Disease: Two Electronic Health Record Prospective Studies'. *The Journal of Clinical Endocrinology & Metabolism* 101 (3): 945–52. https://doi.org/10.1210/jc.2015-3444.

Love, Steve, Sudhanshu Bhatnagar, Greg Rickman & Jedy Wang. 2016. 'The Value of EMR Data: Unlocking Insights That Drive Pharma Sales'. *Journal of the Pharmaceutical Management Science Association*, Spring. As of 11 August 2019: https://www.zs.com/publications/articles/value-of-emr-data-unlocking-insights-that-drive-pharma-sales.aspx.

Lyons, Ronan. 2014. 'Re: Care.Data: Why Are Scotland and Wales Doing It Differently?' *BMJ* 348. As of 11 August 2019: https://www.bmj.com/content/348/bmj.g1702/rr/687637.

Makady, Amr, Renske Ten Ham, Anthonius de Boer, Hans Hillege, Olaf Klungel, & Wim Goettsch. 2017. 'Policies for Use of Real-World Data in Health Technology Assessment (HTA): A Comparative Study of Six HTA Agencies'. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research* 20 (4): 520–32. https://doi.org/10.1016/j.jval.2016.12.003.

Marjanovic, Sonja, Ioana Ghiga, Miaoqing Yang & Anna Knack. 2017. *Understanding Value in Health Data Ecosystems*. Santa Monica, Calif.: RAND Corporation. RR-1972-EFPIA. As of 11 August 2019: https://www.rand.org/pubs/research_reports/RR1972.html.

McCartney, Margaret. 2014. 'Care.Data: Why Are Scotland and Wales Doing It Differently?' *BMJ* 348: g1702. https://doi.org/10.1136/bmj.g1702.

Miani, Céline, Enora Robin, Veronika Horvath, Catriona Manville, Jonathan Cave & Joanna Chataway. 2014. *Health and Healthcare: Assessing the Real-World Data Policy Landscape for Health and Healthcare in Europe*. Santa Monica, Calif.: RAND Corporation. RR-544-PI. https://www.rand.org/pubs/research_reports/RR544.html.

Miettinen, Olli. 1974. 'Confounding and Effect-Modification'. *American Journal of Epidemiology* 100 (5): 350–53. https://doi.org/10.1093/oxfordjournals.aje.a112044.

Nair, Sunil, Douglas Hsu & Leo Anthony Celi. 2016. 'Challenges and Opportunities in Secondary Analyses of Electronic Health Record Data'. In *Secondary Analysis of Electronic Health Records*, 17–26. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_3.

Nelson, Gregory S. 2015. 'Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification'. As of 11 August 2019: http://support.sas.com/resources/papers/proceedings15/1884-2015.pdf.

NICE. 2018. 'Atezolizumab for Treating Locally Advanced or Metastatic Non-Small-Cell Lung Cancer after Chemotherapy. Technology Appraisal Guidance [TA520]'. London: NICE. As of 11 August 2019: https://www.nice.org.uk/guidance/ta520/.

NIH. 2017. 'NIH's Definition of a Clinical Trial'. As of 11 August 2019: https://grants.nih.gov/policy/clinical-trials/definition.htm.

NIH. 2019a. 'List of Registries'. As of 11 August 2019: https://www.nih.gov/health-information/nih-clinical-research-trials-you/list-registries.

NIH. 2019b. 'Fabry Disease'. Genetics Home Reference. As of 11 August 2019: https://ghr.nlm.nih.gov/condition/fabry-disease.

NIHR. n.d. 'Clinical Record Interactive Search (CRIS)'. Accessed 14 August 2019. https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/.

Nordon, Clementine, Constance Battin, Helene Verdoux, Josef Maria Haro, Mark Belger, Lucien Abenhaim & Tjeerd Pieter van Staa. 2017. 'The Use of Random-Effects Models to Identify Health Care Center-Related Characteristics Modifying the Effect of Antipsychotic Drugs'. *Clinical Epidemiology* 9: 689–98. https://doi.org/10.2147/CLEP.S145353.

PARENT. 2015. 'Patient Registries of Europe: Glossary'. As of 11 August 2019: http://parent-ror.eu/#/page/6.

Patel, Rashmi, Nishamali Jayatilleke, Matthew Broadbent, Chin-Kuo Chang, Nadia Foskett, Genevieve Gorrell, Richard D. Hayes, et al. 2015. 'Negative Symptoms in Schizophrenia: A Study in a Large Clinical Sample of Patients Using a Novel Automated Method'. *BMJ Open* 5 (9): e007619. https://doi.org/10.1136/bmjopen-2015-007619.

Petri, Hans, and John Urquhart. 1991. 'Channeling Bias in the Interpretation of Drug Effects'. *Statistics in Medicine* 10 (4): 577–81.

Pijpers, F. n.d. 'Achmea (Agis) Health Database (AHD)'. As of 11 August 2019: http://www.jpi-dataproject.eu/Home/Database/226?topicId=1.

Pushpakom, Sudeep, Francesco Iorio, Patrick A. Eyers, K. Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, et al. 2019. 'Drug Repurposing: Progress, Challenges and Recommendations'. *Nature Reviews Drug Discovery* 18 (1): 41–58. https://doi.org/10.1038/nrd.2018.168.

Roberto, Giuseppe, Ingrid Leal, Naveed Sattar, A. Katrina Loomis, Paul Avillach, Peter Egger, Rients van Wijngaarden, et al. 2016. 'Identifying Cases of Type 2 Diabetes in Heterogeneous Data Sources: Strategy from the EMIF Project'. *PLOS ONE* 11 (8): e0160648. https://doi.org/10.1371/journal.pone.0160648.

Roberts, C. Michael. 2014. 'Chronic Obstructive Pulmonary Disease Audit – Turning Data Into Better Care for Patients'. *Archivos de Bronconeumología (English Edition)* 50 (7): 263–64. https://doi.org/10.1016/j.arbr.2014.05.004.

Roitmann, Eva, Robert Eriksson & Søren Brunak. 2014. 'Patient Stratification and Identification of Adverse Event Correlations in the Space of 1190 Drug Related Adverse Events'. *Frontiers in Physiology* 5. https://doi.org/10.3389/fphys.2014.00332.

Safran, Charles. 2014. 'Reuse of Clinical Data'. *Yearbook of Medical Informatics* 9 (1): 52–4. https://doi.org/10.15265/IY-2014-0013.

Safran, Charles, Meryl Bloomrosen, W. Edward Hammond, Steven Labkoff, Suzanne Markel-Fox, Paul C. Tang & Don E. Detmer. 2007. 'Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper'. *Journal of the American Medical Informatics Association* 14 (1): 1–9. https://doi.org/10.1197/jamia.M2273.

Shire Outcome Surveys. 2017. 'Fabry Outcome Survey. Annual Report 2016. Reporting Period: 17-04-2001 to 05-01-2017'. VV-HEOR-0002. FOS Steering Committee. As of 11 August 2019: http://www.fabrynetwork.org/wp-content/uploads/2017/11/FOS-Patient-Report-2016-Final.pdf.

Singh, Gurparkash, Duane Schulthess, Nigel Hughes, Bart Vannieuwenhuyse & Dipak Kalra. 2018. 'Real World Big Data for Clinical Research and Drug Development'. *Drug Discovery Today* 23 (3): 652–60. https://doi.org/10.1016/j.drudis.2017.12.002.

Sundhed. 2016. 'Background'. sundhed.dk, 13 June 2016. As of 11 August 2019: https://www.sundhed.dk/borger/service/om-sundheddk/ehealth-in-denmark/background/.

The Academy of Medical Sciences. 2016. 'Real World Evidence. Summary of a Joint Meeting Held on 17 September 2015 by the Academy of Medical Sciences and the Association of the British Pharmaceutical Industry'. London: The Academy of Medical Sciences. As of 11 August 2019: https://acmedsci.ac.uk/file-download/38667-573d8796ceb99.pdf.

The Academy of Medical Sciences. 2018. 'Next Steps for Using Real World Evidence: Summary Report of a FORUM Follow-up Roundtable Held on 24 January 2018'. As of 11 August 2019: https://acmedsci.ac.uk/file-download/7021031.

The London School of Economics and Political Science. 2018. 'RWE in Europe Paper V: Policy Challenges around Real World Evidence Adoption in Europe'. December 2018. As of 11 August 2019: http://www.lse.ac.uk/business-and-consultancy/consulting/assets/documents/rwe-in-europe-paper-v.pdf.

The World Health Organisation. n.d. 'International Clinical Trials Registry Platform Search Portal'. As of 11 August 2019: http://apps.who.int/trialsearch/.

Tikkanen, Roosa. 2017. 'Multinational Comparisons of Health Systems Data, 2017'. 15 November 2017. As of 11 August 2019: http://www.commonwealthfund.org/publications/chartbooks/2017/multinational-comparisons-2017.

UCL Institute of Epidemiology & Health Care. n.d. 'THIN Database Research Group'. As of 11 August 2019: https://www.ucl.ac.uk/epidemiology-health-care/research/primary-care-and-population-health/research/thin-database/database.

US FDA. 2018. 'Use of Electronic Health Record Data in Clinical Investigations Guidance for Industry'. 83 FR 34137. Silver Spring, Maryland: FDA. As of 11 August 2019: https://www.federalregister.gov/documents/2018/07/19/2018-15390/use-of-electronic-health-record-data-in-clinical-investigations-guidance-for-industry-availability.

Virtual Trials Archives (homepage). n.d. As of 11 August 2019: http://www.virtualtrialsarchives.org/.

Virtual Trials Archives. n.d. 'VISTA Data Request Form'. As of 11 August 2019: http://www.virtualtrialsarchives.org/data-request-form/.

VISTA. n.d. 'The Virtual International Stroke Trials Archive'. As of 11 August 2019: http://www.virtualtrialsarchives.org/vista/.

Wang, Jian. 2011. 'Data Exploration in Secondary Use of Healthcare Data'. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, 658–658. https://doi.org/10.1109/BIBM.2011.129.

Wise, John, Alexandra Grebe de Barron, Andrea Splendiani, Beeta Balali-Mood, Drashtti Vasant, Eric Little, Gaspare Mellino, et al. 2019. 'Implementation and Relevance of FAIR Data Principles in Biopharmaceutical R&D'. *Drug Discovery Today*, January. https://doi.org/10.1016/j.drudis.2019.01.008.

Wise, John, Angeli Möller, David Christie, Dipak Kalra, Elia Brodsky, Evelina Georgieva, Greg Jones, et al. 2018. 'The Positive Impacts of Real-World Data on the Challenges Facing the Evolution of Biopharma'. *Drug Discovery Today* 23 (4): 788–801. https://doi.org/10.1016/j.drudis.2018.01.034.

Yildirim, Oktay, Matthias Gottwald, Peter Schüler & Martin C. Michel. 2016. 'Opportunities and Challenges for Drug Development: Public–Private Partnerships, Adaptive Designs and Big Data'. *Frontiers in Pharmacology* 7. https://doi.org/10.3389/fphar.2016.00461.

Annex A. Methodological approach

A.1. Overview of methodological approach

The objectives of this study were to understand:

(i) How different types of health data are reused by the pharmaceutical industry (e.g. what are they being used for) and reasons for this.
(ii) The key enablers and barriers to effective reuse of data.
(iii) Considerations and potential implications for future action by different stakeholders (including industry and policymakers).

In order to meet these objectives, we carried out a targeted literature review, identified and developed a set of case vignettes, ran a series of stakeholder interviews and held a synthesis workshop. More details on each of these elements are provided below.

A.2. Targeted literature review

A rapid review of the literature was conducted to gain an understanding of the types of health data the pharmaceutical industry may use, how the industry uses this data, the enablers and challenges of reusing health data and the possible implications for pharmaceutical companies and beyond.

The search of Google Scholar (covering articles published between 2009 and 2018 inclusive) included the following search terms:

> "Prescription data" OR "electronic medical records" OR EMR OR "electronic health records" OR EHR OR "electronic patient records" OR "EPR" OR "health registr*" OR "health systems data" AND "use" OR "reuse" AND "pharma* industry"

For a document to be included in this review, it had to include the following:

- Secondary use of health data (i.e. using data for something other than it was originally collected for).
- Use of data by the pharmaceutical industry: publications were defined as being pharmaceutical industry publications if at least one of the authors had a pharmaceutical industry affiliation.[15]

[15] In practice, this can mean a variety of ways of how pharmaceutical industry stakeholders were involved which are not necessarily exclusive: the pharmaceutical company itself is doing the analyses (possibly with other stakeholders); the

- Use in Europe.
- Focus on current examples of secondary use of health data, rather than prospective use.

A total of 23 academic papers and grey literature documents were included following the criteria above. Relevant information from these documents was extracted. This included the country(-ies) of focus, the purpose of the document/aim of the research, what type of health data was used, why the data was used, how the data was used, enablers, challenges and implications for future secondary use of data. Information extracted through these 23 documents was used to snowball and identify additional relevant literature for the literature review and to identify potential case vignettes. In total, 41 documents were analysed for their relevance and useful data extracted where possible (although not all of these were included in the literature review).

It is important to highlight that there is a wide range of published literature on the secondary use of health data, but the majority either focuses on academia's reuse of these data or the context is outside Europe.

A.3. Case vignettes

We used a mixed-methods approach to collect and analyse data to develop 12 case vignettes to illustrate how the European pharmaceutical industry currently reuses health data (Annex A).

A.3.1. Identification of examples for the case vignettes

Three main sources were used to identify potential case vignettes: targeted literature searches (Section A.1), snowballing from literature searches and three initial key informant interviews with pharmaceutical industry representatives.

In a first step, we tried to find interesting case examples through our literature review (Section A.1) and through targeted searches of Google Scholar and Google. Targeted searches included:

- Search of different types of health data (e.g. EHR, registry data, prescription data) in conjunction with the names of different European pharmaceutical companies.
- Search of commonly used datasets (e.g. Hospital Episode Statistics (HES), The Health Improvement Network (THIN), Clinical Practice Research Datalink (CPRD)) in conjunction with the names of different European pharmaceutical companies.

Literature was eligible for inclusion if a representative of a European pharmaceutical company was listed as an author of the publication. We used a snowballing approach to identify additional potentially relevant literature from the reference lists of the initially found publications. In three interviews with representatives of pharmaceutical companies, we explored further opportunities for case vignettes.

---

pharmaceutical company is funding the study and the analyses are being conducted by collaborators; the pharmaceutical industry is providing steering for the project and the analyses are being conducted by collaborators; the pharmaceutical company is providing the data and support for a study conducted by others.

For each identified study, we extracted key information into an initial extraction template, which included the following:

- Bibliographic information
- Pharmaceutical company/-ies involved in the study
- Country/-ies where the study was conducted[16]
- The type(s) of data reused
- Purpose(s) of reusing the data
- Focus of interest of the study (e.g. disease area).

A.3.2. Selection of the case vignettes

We identified 12 case vignettes which satisfied the following criteria:

- Available literature provides sufficient evidence required to describe the secondary use of health data.
- A balance of different data types.
- A balance of different purposes of reusing the data.
- Focus of interest of the study (e.g. disease area).

These were presented to EFPIA and the study steering group. All 12 case vignettes were considered eligible for a further exploration and included in this study.

A.3.3. Development of the case vignettes

We further developed the extraction template used at the case vignette identification stage and extracted the following additional information:

- Aim of the study
- How data was used
- Enables of reusing the data
- Challenges to reusing the data
- Implications for future research, industry or policy
- Current/prospective use (i.e. whether the article discusses the current or future use of health data)
- Other relevant notes.

The extraction template is presented below.

[16] i.e. the articles' authors' organisations' or companies' location listed in the publications.

Table 2: Case vignette extraction template

Category
Bibliographic information
Pharmaceutical company/-ies involved
Country/-ies where the study was conducted
Aim of the study
Type(s) of data reused
Purpose(s) of reusing the data
How data was used
Enablers of reusing data
Challenges to reusing data
Implications for future research, industry or policy
Focus of interest of the study (e.g. disease area)
Current/prospective use (i.e. whether the article discusses the current or future use of health data)
Other relevant notes

For each selected study, we invited the pharmaceutical industry authors to an interview to complement information presented in the publications, get deeper insights as well as to clarify and fill any gaps related to the reuse of data which were not discussed in the literature. The interviews were semi-structured; while each interviewee was asked the same set of questions related to the items of the extraction template, we also allowed for other issues to be explored as they emerged. Before conducting the interview, each interviewee was provided with an overview of the project and was asked for their consent regarding recording, using and storing information collected (in line with General Data Protection Regulation (GDPR) 2018).

Interviews were conducted by telephone and lasted between 20 and 50 minutes. A total of seven case vignette interviews were conducted, covering 5 of the 12 case vignettes.[17]

The information extracted as well as insights from interviews were used to write up the case vignettes, which followed a standardised template to allow for comparability across case vignettes (see Table 3).

Table 3: Case vignette template

Case vignette title
Summary box • A box summarising the key findings of the case vignette. **Who was involved in the study?** • A summary the different stakeholders who were involved in the study.

[17] For the remaining seven vignettes, we either did not get a response or the interview was declined.

Case vignette title
What were the aims? • A short description of the aims of the study and how the secondary use of health data should support achieving these. **What data were used?** • A short description of the types of data used in the study. **How and why did they use the data?** • A summary of how data were used by industry and what the purpose of using these data was. **What were the enablers and challenges of the study?** • A description of what helped conduct the study and what barriers were observed, with a particular focus on the enablers and challenges related to the reuse of health data. **What safeguards were employed to govern the use of the data?** • A description of how the participating organisations ensured the safe use of the data. **What were the potential or realised benefits to patients and public health?** • A description of any benefits to the patients and the public health this study has or may have realised.

A.4. Stakeholder interviews

In addition to the case vignettes, we conducted key informant interviews with representatives of European pharmaceutical companies and regulatory authorities to better understand the current secondary use of health data. Similar to the case vignette interviews, we used a semi-structured interview protocol with the same set of questions asked in each interview, each interviewee provided their consent in line with GDPR 2018 prior to the interview, interviews were audio recorded and detailed notes of the conversations were taken. The telephone interviews lasted between 20 and 60 minutes.

Potential interviewees were identified in the reviewed literature from our own professional networks through targeted Google searches for representatives of the European pharmaceutical industry, and through suggestions made by EFPIA and by interviewees. A total of ten interviews (eight pharmaceutical industry and two regulatory authority representatives)[18] were conducted; two case vignette interviewees were also asked the same set of questions in addition to the case vignette-specific ones.

Interview notes were transferred into an analysis template which followed the themes explored during the interviews:

- Types of health data being used by the pharmaceutical industry in Europe.
- How the European pharmaceutical industry is currently using use these different types of health data.
- The reasons why the European pharmaceutical industry is reusing health data.

[18] INT4, INT9

- Enablers and barriers associated with the (effective) reuse of health data.
- Positive and negative impacts in relation to the reuse of health data by the European pharmaceutical industry.
- Perceptions about what will happen with the secondary use of health data by the European pharmaceutical industry in the near future.
- Particular things that need to be considered to get the most out of health data in terms of implications for future research, industry or policy.

Interview responses relating to each of these sections were analysed and narratively synthesised with findings from the literature and the case vignettes. Interview inputs are cited within the report using the code 'INT' immediately followed by a number between 1 and 16 (e.g. INT1, INT2, and so on), and references to the case vignettes are cited using the code 'CV' immediately followed by number between 1 and 12, which refer to the case vignette numbers provided in Chapter 2 (e.g. CV1, CV2, and so on).[19]

A.5. Synthesis workshop

In the final phase of the research, we organised a workshop with EFPIA and members of the EFPIA steering group for the study, which was held on 15 May 2019. At this workshop, we focused on the key findings outlined in Chapters 1 and 4, aiming to 'stress test' the analyses with experts from the pharmaceutical industry as well as to draw out related key implications for future actions by stakeholders within the wider health data ecosystem. Specifically, members of the study team presented short key findings of each section of the chapters, followed by a group discussion focusing in particular on any lessons that could be learnt for the future. A member of the study team captured the findings and future implications on Post-its on a whiteboard, and another member wrote detailed minutes during the discussion. Following the workshop, the study team used the reflections covered on the Post-its and in the minutes to refine and strengthen Chapters 1 and 4 as well as to provide considerations for the future (presented in Chapter 5).

A.6. Caveats of the analysis

There are a few caveats that should be borne in mind when interpreting the findings presented in this report:

- While the literature review conducted for this study followed a structured search approach, searches were constrained by a few factors (e.g. database searched, search terms, only including literature published in English) and therefore some relevant publications may have not been identified.

[19] In the case vignettes, in order to retain anonymity, we do not use interviewee identifiers. In the introduction to the case vignettes in which use interview data, we have noted that an interview has contributed to the vignette.

- As highlighted by several interviewees, there has been a rapid increase of the reuse of health data by the pharmaceutical industry recently, but much of what industry has done so far has not been published yet (or is not in the public domain). As a result, there may be relevant cases of the reuse of health data by the European pharmaceutical industry, but these could not be captured by our literature searches. To mitigate this caveat, we engaged with pharmaceutical industry representatives in interviews.
- Although we tried to speak to at least one industry representative per case vignette, we were unable to get a positive response from each invited individual, and hence each case vignette.
- The focus of this study was on the reuse of health data by the European pharmaceutical industry. We are aware that most large pharmaceutical companies operate in several countries and on several continents. As we only included articles where a specific focus on Europe (or a European country) was explicitly mentioned, and as we only selected articles as case vignettes where industry authors were explicitly stated to be located in a European country, we may have missed relevant examples of the European pharmaceutical industry reusing health data.
- This report was not intended to be a comprehensive overview of *all* secondary analysis activities undertaken by the European pharmaceutical industry, but should rather be a snapshot view of key developments in the reuse of health data by industry, key enablers and challenges observed, and aimed to identify what needs to be done to improve the reuse of health data by European industry.
- Related to the aim to develop a snapshot view of the European pharmaceutical industry reusing health data, the case vignette descriptions in this report are relatively short and do not include all available literature as well as details presented in the reviewed publications. The purpose was to provide a brief overview and to illustrate the current reuse of health data.