Understanding value in health data ecosystems

A review of current evidence and ways forward

Sonja Marjanovic*, Ioana Ghiga*, Miaoqing Yang, Anna Knack

*Corresponding authors:
Dr Sonja Marjanovic (Research Leader, RAND Europe): smarjano@rand.org
Ioana Ghiga (Analyst and project manager, RAND Europe): ighiga@rand.org
Acknowledgements and Contact Details

The authors would like to thank our quality assurance reviewers, Jon Sussex and Dr Advait Deshpande, for their constructive guidance through their quality assurance roles. We would also like to acknowledge all the workshop participants and interviewees who gave us their time and who provided valuable information for this project, as well as all the individuals we engaged with as part of the project working group at the European Federation of Pharmaceutical Industry Associations (EFPIA), and in particular Ludivine Blanc, Dr Shahid Hanif and Dr Brendan Barnes.

For more information about this report, please contact:

Dr Sonja Marjanovic
Research Leader, RAND Europe
Westbrook Centre, Milton Road
Cambridge, CB41YG
smarjano@rand.org
A need to understand changing data environments in the health sector

The evidence base that informs the life sciences and healthcare delivery is changing substantially. New types of data are informing the health sector, and advances in digital technologies are allowing for wider-scale and more diverse data gathering, processing and analysis. In its broadest sense, health data refers to any type of data that is useful for improved research, innovation and healthcare-related decision making. It can inform research and innovation, prevention and treatment strategies, health promotion efforts, self-care, health systems planning and wider public health activities, behaviours and decisions. As such, health data is relevant to a broad range of actors: academic researchers, industry, healthcare professionals and providers, patients and the public, payers and policymakers, regulators, charities and the third sector.

Health data stems from diverse sources and includes clinical trials data, clinician- or patient-reported data, and behavioural and health systems data. Some examples include: (i) electronic health records (EHR) data on patient symptoms, referrals, prescriptions and treatment outcomes; (ii) longer-term treatment outcomes data from real-world effectiveness studies; (iii) medicine performance data from randomised controlled trials (RCTs); (iv) genomic and proteomic data on individuals and associated biomarker data; (v) data from wearables and sensors (e.g. on vital signs); and (vi) data on individual preferences and health-seeking behaviours from social media. Although there is no consistent or unified definition of health data in the literature, it tends to be categorised according to who the providers are or what the data attributes or methods of collection are. In this context, literature devotes substantial attention to the potential of big data and real-world data to inform the health sector.1

In light of the changing evidence base underpinning the health sector, the European Federation of Pharmaceutical Industry Associations (EFPIA) commissioned RAND Europe2 to conduct a rapid evidence review of key insights on the value of health data. More specifically, the review aimed to: (i) identify and explain the potential and existing benefits that can stem from effective use of health data; and (ii) examine the key drivers of supportive health data ecosystems and their implications for future research, policy or practice, in light of the diverse challenges to be addressed. By health data ecosystems, we mean the technological and social arrangements underpinning the environments in which health data is generated, analysed, shared and used. The review was informed by an analysis of scholarly and grey literature, interviews with experts from diverse stakeholder groups and consultation at a European Union (EU)-level expert roundtable.

---

1 Big data is generally defined based on attributes of volume, velocity of data processing, variety, veracity and value. Real-world data refers to data collected (prospectively or retrospectively) outside the context of traditional RCTs and from diverse sources, for example data from EHR, observational media.

2 RAND Europe is a not-for-profit policy research organisation that helps to improve policy and decision making in the public interest, through research and analysis. RAND Europe’s clients include European governments, institutions, NGOs and firms with a need for rigorous, independent, multidisciplinary analysis. This document has been peer reviewed in accordance with RAND Europe’s quality assurance standards and as such can be portrayed as a RAND Europe document.
The potential benefits of health data access, sharing and use

The literature on health data generation and use identifies a myriad of potential social and economic benefits from health data use, as overviewsed in Table 1. Overview of potential benefits of health data In general, the evidence base focuses more on future than realised potential, reflecting the state of the field. Throughout the report we provide examples of initiatives seeking to capture value in three main areas: (i) research and development (R&D) and innovation; (ii) public health and pharmacovigilance; and (iii) healthcare delivery and the health system more widely. We also discuss how specific stakeholder groups could benefit from a data-rich health ecosystem.

Table 1. Overview of potential benefits of health data

<table>
<thead>
<tr>
<th>Potential benefits for R&amp;D and innovation</th>
<th>Potential benefits for public health and pharmacovigilance</th>
<th>Potential benefits for healthcare delivery and the wider health system</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Opportunities to explore new research areas could stem from access to richer data sources and new analysis techniques. For example: (i) linking datasets on genetic profiles with EHR on patient symptoms can help reveal patterns of association or disease causation that were previously not possible to detect; (ii) access to real-world data from pragmatic trials and other real-world evidence can enable research not possible under an RCT model due to ethical issues (e.g. studies on narcotic abuse) or due to challenges of sample size (e.g. research on rare diseases).</td>
<td>• Prospects to scale up use of real-world health data in pharmacovigilance: Access to more diverse and greater amounts of real-world data than currently practised in pharmacovigilance (drug-safety monitoring), coupled with more granular information on patient profiles, could facilitate quicker and more rigorous learning about how drug safety relates to particular patient groups over time, including in the context of co-morbidities.</td>
<td>• Benefits for healthcare quality: For example: (i) more personalised care and enhanced predictive analytics could be enabled by more comprehensive clinical datasets (e.g. improved screening algorithms and integration of imaging data, genomic and proteomic data on new biomarkers, and symptoms data from EHR); (ii) workforce access to more comprehensive evidence could facilitate better-informed care decisions (provided that evidence is presented in a user-friendly manner and trusted).</td>
</tr>
<tr>
<td>• Operational and cost-efficiencies could stem from: (i) better targeting of R&amp;D investments and more appropriate clinical trial design due to improved patient stratification based on genetic traits and clinical records; (ii) reduced unnecessary duplication in research and enhanced confidence in results due to access to a richer and broader evidence base and enhanced data sharing.</td>
<td>• Prospects for enhanced, data-enabled public health promotion and prevention strategies: For example: (i) large and integrated environmental, genetic and socio-economic datasets could enable better prediction of risk factors for disease; (ii) data on health apps and portable devices could enable citizen empowerment and proactive behaviours in maintaining good health; (iii) computer algorithms and predictive analytics could assist in disease screening and early diagnosis.</td>
<td>• Operational and cost-efficiencies in healthcare delivery: For example: (i) easier comparability of outcomes data from different treatments across patient profiles could be enabled by large datasets from EHR, and could allow for more efficient decision-making, reducing wastage and costs associated with administration of inappropriate or inferior treatments; (ii) a reduction in unnecessary hospitalisations could be facilitated through data- and technology-enabled self-care and self-management of risk factors and through remote monitoring of adherence to treatments (this would require careful risk management).</td>
</tr>
<tr>
<td>• Health data can also enhance the quality of research and innovation processes and outputs. For example: (i) real-world data from pragmatic trials can increase confidence in study results given that sample populations may be more representative of actual practice; (ii) using real-world data throughout the R&amp;D and innovation cycle could also facilitate reimbursement for products that have a proven enhanced efficacy in a real-life setting and inform value-based and outcome-based payment approaches, as well as adaptive pathways; (iii) longitudinal data on treatment adherence and compliance creates prospects for new outcome measures in research.</td>
<td>• Emergency-preparedness could be improved through more timely data matching disease outbreaks with covariates (such as environmental data from satellite sensors and data on symptoms from both health professionals and social media (although checks on reliability would be needed).</td>
<td>• Wider benefits for the health system: For example: (i) real-world outcomes data for treatments (e.g. clinical and patient experience data) could enable better-informed drug safety regulation, adaptive pathways and innovative reimbursement models; (ii) greater usage of EHR and costs data could facilitate more efficient health systems planning and resourcing, improved workflows and administration efficiency.</td>
</tr>
</tbody>
</table>
Creating supportive health data ecosystems

The landscape for health data generation, interpretation and use is still in a relatively early phase of evolution, with potential for much further growth. A supportive health data ecosystem, which would harness the full potential of health data and scale up existing benefits, is yet to be established. To achieve this, further research and policy consideration in a range of areas is required (as highlighted in Figure 1. Building blocks of supportive health data ecosystems) – reflecting both the opportunities and the inevitable challenges that accompany transformational initiatives of the scale presented by a health data-rich society. We discuss each of these key areas below.

Figure 1. Building blocks of supportive health data ecosystems

Collaboration and coordination in a health data ecosystem: Initiatives which seek to capture value from health data need to take account of the interdependencies and interactions between both stakeholders (e.g. research, industry, healthcare providers, patients, policy) and sectors (e.g. health and social care), given that health R&D and the delivery of healthcare are increasingly cross-sectoral activities. There is much more focus in the literature on potential benefits for a particular stakeholder group than on how value capture by any one group depends on its interactions with others. Similarly, there is little focus on how collective collaboration and coordination could work to address key structural, technical, legal and social boundaries to data sharing and access – including ownership and economic issues. In this context, the alignment of interests and the effective management of trade-offs is critical and individuals and organisations with boundary-spanning roles across different communities will be instrumental in addressing barriers and scaling up existing efforts.

Public acceptability and engagement: Literature highlights a need for learning from prior efforts to inform future engagement and communication campaigns with the public and to enhance public acceptability and involvement with health data sharing and use. The evidence emphasises that patients and the public expect to be informed. They need to understand the opportunities, risks and safeguards associated with health data use, individual and wider public-good benefits, potential impacts on doctor–patient relationships, and how inequalities will be prevented. The public expects to be informed about who has access to which data and why, how sensitive data and data portability will be governed and managed, under what circumstances they may opt out, and what legal provisions and citizen rights are in place should unintended uses or data hacking take place.
Future research needs to consider what trade-offs patients and the public are willing to make when it comes to managing benefits and risks.

Making the most of recent data protection regulation and considering data sharing models: Developments such as the EU’s General Data Protection Regulation (GDPR) are paving the way for a clearer governance framework for anonymised health data use in research and care delivery. The GDPR seeks to provide simplified consent requirements and requires rigorous technical and organisational safeguards for data sharing and use. Although the GDPR is a promising development in many ways, it is not without its challenges. For example, the boundaries of what is considered research use (and what constitutes commercial activity) and associated eligibility criteria for access to data are not yet clear. How the GDPR will be implemented at both EU and national levels remains to be seen. There is a need for further clarity on how different EU member states will interpret and act on the guidance, and what this implies for markets for health and analytics, going forward. An active campaign to ensure public understanding and acceptability of the GDPR governance framework will also be important. Further research will be needed to understand the regulatory capacity that has to be in place to ensure sufficient oversight of data access and sharing practices, public trust and commercial confidence in data validity. Related to data access and data sharing practices (but not the GDPR exclusively), various privacy and security safeguards are being piloted globally and span technological solutions (e.g. anonymisation, block chain, bloom filters) and more social and institutional responses (e.g. the GDPR, dynamic consent models, trusted curators).

Data quality and technical considerations: The literature on health data access and use also discusses a series of technical issues that initiatives in the health data field face. These include issues related to data quality assurance, a need for harmonised standards, issues related to the compatibility of IT infrastructure, a need for secure forms of data storage and transfer and enhanced data aggregation and sense-making capacities, and the governance of data use through technology and social interventions. There is already a diversity of EU efforts under way to try to address these aspects of health data ecosystems. Some examples include various cloud computing efforts for data storage solutions, new data mining techniques such as machine learning, opportunities for producing more complex databases with more complex data relationships, techniques for speedier data processing, and new probabilistic matching techniques to manage challenges associated with data linkage and anonymisation.

Workforce skills and capacities to engage with health data: Improving the ability of healthcare professionals to engage with health data will require capacity-building in both technical skills (e.g. clinical informatics) and in softer skills related to leadership and communication of new sources and types of evidence. Healthcare professionals will need to know how to access, interpret and act on diverse types and vast amounts of health data. Capacity-building among the workforce will also require the addressing of: concerns related to liability associated with data use and decision-making; time demands associated with engagement; uncertainty around how the doctor–patient relationship could be affected; and wider workforce implications. Capability-building needs to be considered in educational curriculums and continual professional development programmes alike. The research community, regulatory bodies and the general public also need to strengthen capacity to engage with health data in education and communication activities.
In reflection

The potential of health data to improve the efficiency and effectiveness of health R&D, healthcare delivery and health systems more widely is substantial. There are many initiatives across the EU that are experimenting with ways to capture value and address the nexus of technical, legal, ethics-related, governance and data protection-related and cultural challenges to delivering potential benefits for society and the economy. The field of health data research and policy is highly dynamic and there is a need for further reflection, thematic learning and evaluation to better understand how to create and connect receptive places, to inform future interventions and to identify transferable lessons. Our research emphasises that realising the benefits of health data at scale will require: a simultaneous focus on the technological and structural conditions that are required; collaboration and coordination to transform working cultures and build health and care workforce and citizen capacity to engage with data; and efforts to ensure that policy, industry and research communities respond to public concerns, needs and expectations in a timely and sustained manner. The global community of individuals and organisations with a stake in health data will also need to consider how progress can benefit different populations across the world, in an equitable manner.
# Table of Contents

Acknowledgements and Contact Details ........................................................................... i
Summary ............................................................................................................................... ii
Table of Contents ............................................................................................................... vii

1. **Background and context ........................................................................................... 1**

2. **Key types of health data .......................................................................................... 2**

3. **Different uses of health data and associated value .................................................. 5**
   3.1. Potential benefits for the R&D process ................................................................. 5
   3.2. Benefits for pharmacovigilance and public health .............................................. 10
   3.3. Benefits for healthcare delivery and the health ecosystem more widely .......... 14
   3.4. Who could benefit from health data? ................................................................. 19

4. **Enabling receptive health data ecosystems .............................................................. 21**
   4.1. Considering the value propositions presented by health data in the context of interdependencies and interactions between stakeholders and sectors .......... 22
   4.2. Building on existing efforts: momentum to address data quality and technical considerations ................................................................. 24
   4.3. Making the most of recent data protection advancements and considering models of data sharing ................................................................. 25
   4.4. Workforce capacity-building: supporting health professionals to make the most of health data ......................................................................... 28
   4.5. Learning from previous efforts to increase public awareness, acceptability and engagement with health data ................................................................. 28
   4.6. In reflection ........................................................................................................... 30

References .......................................................................................................................... 32

Appendix A Methodological Approach .............................................................................. I
Appendix B List of Consulted Experts ................................................................................ III
Appendix C Value of Health Data – Selected Examples ..................................................... IV
1. Background and context

The evidence base for the life-sciences sector and for healthcare delivery is changing. New types of health data are presenting new opportunities for research and innovation, healthcare delivery and health systems more widely, by creating channels for better-informed decision-making and behavioural change. As discussed in sections 2 and 3 of this report, although health data has become an umbrella term, it can – in its broadest sense – refer to any type of data that is useful for improved research, innovation and healthcare-related decision-making – at individual, organisational and system levels. Diverse sources of health data (e.g. clinical trial data, clinician- or patient-reported data, behavioural and health systems data) are opening new prospects for improving population health through more personalised care and more empowered and engaged patients and citizens, better-quality and timelier healthcare services, and new diagnostic and treatment possibilities (Miani et al. 2014). Research and innovation landscapes are becoming ever more open and collaborative with new stakeholders entering the sector, in part enabled by internet-based technologies and advances in data science that allow for new means of data collection and analysis (Martin-Sanchez & Verspoor 2014; Miani et al. 2014). In this changing landscape, there is a need for researchers, industry, healthcare professionals, patients and the public, regulators, payers and policymakers to consider a richer and broader array of evidence in decision making, within a supportive and sustainable health-data ecosystem.

To better understand the value proposition (i.e. potential for benefit or already-realised benefits) of different types of health data, the European Federation of Pharmaceutical Industry Associations (EFPIA) commissioned RAND Europe to conduct a rapid evidence review of current insights on the topic. More specifically, the review aimed to analyse insights from the current knowledge base to: (i) identify and explain the potential and existing benefits of the effective use of health data for research and development (R&D) and innovation, healthcare delivery and the functioning of healthcare systems more widely; and (ii) examine the required elements of a supportive data ecosystem and their implications for future research, policy and practice.

The contents below discuss our key findings. Section 2 overviews the diversity of health data types discussed in scholarly and grey literature. Section 3 builds on this understanding to consider the diversity of benefits that could be realised from health data generation and use. Section 4 reflects on the literature and our consultations with experts to discuss critical areas for future consideration in efforts to establish supportive health data ecosystems.1 Advancing beyond the relatively fragmented landscape for health data use at present, and capturing value from health data at pace and at scale, will depend on the ability of stakeholders to address a diversity of structural, technological and quality-related, behavioural, cultural and incentive-related challenges. These have been discussed in diverse literature (e.g. Alyass et al. 2015; European Commission [EC] 2016; Geissbuhler et al. 2013; Hoffman & Podgurski 2013; Holtorf et al. 2008; Schneeweiss 2014; van Panhuis et al. 2014;) and are not the core focus of this report. Rather, we focus on what the literature teaches us about the types of progress needed to create facilitative health data ecosystems, which can effectively navigate and manage inherent tensions and challenges.

---

1 By data ecosystems, we refer to the technological and institutional arrangements underpinning environments in which health data is generated, analysed and used.
The insights discussed in this report primarily draw from the analysis of scholarly and grey literature on the topic from 149 articles, reports and relevant web pages. Out of these, 57 articles and reports were subjected to systematic data extraction after full text review, and the remainder identified through a snowballing approach and analysed for additional and complementary thematic insights. Findings from the literature were enriched with insights from 12 key informant interviews, and views from 22 participants spanning different stakeholder groups, at a project roundtable discussion held on 17 February 2017 (in London, United Kingdom [UK]). Please see Appendix A for more detail on the methodological approach. Appendix B provides a list of individuals consulted as part of the study, either as interviewees or as roundtable participants (where permission to name them was granted).

There are some caveats to consider when interpreting the information in this report. First, although the authors reviewed a diverse range of literature, we are aware that the project did not consider all the available literature on health data that is in the public domain (i.e. was predominantly based on a rapid evidence review which also did not consider initiatives outside of the EU). However, the research design and analysis (see Appendix A) paints a rounded picture of the diversity of potential benefits and the state of the field in terms of capturing value. Second, the review was focused on thematic analysis, and an examination of national and regional differences was not in scope. However, the authors fully recognise that such differences (e.g. in trust in institutions to maintain data security and effectively govern data use) can influence future prospects for capturing value. Third, and as will be apparent in the contents that follow, the literature more frequently discusses future potential than existing and captured value, reflecting current social, institutional and technical phases of development of health data ecosystems and the associated challenges. Finally, there are unanswered questions and future research needs (e.g. in areas such as evaluation, understanding interdependencies between stakeholders and sectors in health data value chains, examining national contextual specificities). We hope that this review will be useful for future studies aiming to advance the field and inform health data ecosystems.

2. Key types of health data

Understanding the health data ecosystem and the transformative potential of health data starts with a need to identify the different types of data which offer value for distinct stakeholders and spheres of health-sector activity and performance. The health data landscape (and associated literature) is currently characterised by a significant plurality of data types (Figure 2) and multiple categorisations. There is a lack of a standardised analytical framework or typology of health data. However, in its widest sense, health data refers to any health-related information that is of relevance to decision-making in a health system and that can inform prevention, treatment, cure, health promotion, self-care and wider public health activities and decisions taken by individuals, families, communities, and policy-makers.

---

4 The 12 interviews have the following distribution: 3 representatives of Academia, Research Institutes, Research Councils and Think Tanks, 3 of Consumer and Patients Associations, 3 of Policy/Government/EC, 3 of Private Sector, Trade Association or Regulatory or Other.

5 The 22 roundtable participants have the following distribution: 2 representatives of Academia, Research Institutes, Research Councils and Think Tanks, 4 of Consumer and Patients Associations, 5 of Policy/Government/EC, 11 of Private Sector, Trade Association or Regulatory or Other.

6 Value chains help describe and analyse the diverse set of activities that are needed to bring a product from conception to intended use and market, and to reinvention and reuse. As described in Parks et al. (forthcoming), a value chain is made up of value-adding ‘links’ between different stages of production. Each link includes a variety of activities that go well beyond simple production, such as the development and design of a product or process, activities related to marketing, and the consumption, as well as recycling, of product materials. By extension this would also imply activities related to aggregation, synthesis, translation, analysis, use and reuse of data.
stakeholders. As such, it is relevant to a broad range of actors, spanning researchers, industry, healthcare professionals, patients and the public, regulators, payers and policymakers. To illustrate, some examples include: data from electronic health records (EHR) on patient symptoms, prescriptions, referrals and treatment outcomes; data on drug side-effects; monitoring data on vital signs from wearable devices and sensors; hospital admissions and performance data; healthcare claims data and data on costs of treatments or services; epidemiological data and disease outbreaks information; scientific data such as results from biomedical studies, imaging data to aid in diagnosis and data from clinical trials; and data on individual preferences and health-related behaviours from social media which can inform health promotion campaigns.

Despite the inherent diversity and complexity entailed in the term ‘health data’, it is possible to identify some key data types based on attributes, sources and methods of collection, as elaborated on below.

Figure 2. The plurality of data types characterising the health data field

In recent years, much attention has been given to big data across all sectors, including health. Although there is no universally accepted and harmonised definition for this concept, several attributes of datasets are used to define big data in the literature. These include volume, velocity, variety, veracity and value – otherwise known as ‘the 5Vs’ (EC 2016; Salas-Vega et al. 2015; EC 2014a; EC 2014b). These features illustrate the large amount of data, the speed at which it is being generated, the different sources and types of data as well as the benefits and transformational capabilities that big data could bring to the health system. Most recently an EC (2016) report proposed the following definition for big data in health, explicitly highlighting additional features such as multi-purpose use and scope for re-use: “Big Data in Health refers to large routinely or automatically collected datasets, which are electronically captured and stored. It is reusable in the sense of multipurpose data and comprises the fusion and connection of existing databases for the purpose of improving health and health system performance.” (22)

Literature on health data specifically highlights the relevance of multiple types of big data (EC 2014a; EC 2014b, EC 2016; INT01–10; Miani 2014; Salas-Vega et al. 2015), including: (i) EHR, which can contain information on symptoms, medical exams, tests, referral patterns, prescriptions and death records as well as pharmacy records, diagnostic procedures, hospitalisations and other healthcare services; (ii) claims data giving indications of the nature of service usage, insurance and other administrative hospital data; (iii) omics data: genomics,

---

7 Throughout the report interviewees are referenced by number, e.g. interviewee 1 is referred to as INT01. Their stakeholder affiliation is not disclosed to ensure anonymity.
transcriptomics, proteomics, epigenomics, metagenomics, metabolomics, nutriomics; (iv) clinical trials data; (v) pharmaceutical data such as pharmacovigilance (medicines safety) data; (vi) social media including web data pertaining to health such as data from patients forums on health topics; (vii) mobile apps, telemedicine and sensor data; (viii) geospatial health data (health data disaggregated by location); (ix) ambient data from ‘smart’ environments (e.g. electricity and gates data on the way people walk which can be used to estimate the occurrence of falling [INT03]); (x) information on well-being, socio-economic, behavioural data; and (xi) other records of relevance to health such as occupational records, sociodemographic profiles or environmental monitoring data such as on pollution.

Herland et al. (2014), in their review of data mining using big data in health informatics, draw similarities between the terms ‘big data’ and health informatics data, based on shared ‘5V’ features. Within this, key subfields of health informatics data discussed span: bioinformatics, which uses molecular-level data; image informatics, which uses tissue-level data such as brain image data in neuroinformatics; clinical informatics, which uses patient-level data; public health informatics and pharmacovigilance, which uses population-level data (including via social media); and translational bioinformatics (TBI), focused on integrating multiple types of informatics data towards improving healthcare outcomes.

Closely associated to big data is the growing attention to real-world data applications in healthcare (Miani et al. 2014). This refers to health data collected outside the context of traditional randomised controlled trials (RCTs) and can stem from diverse sources (e.g. clinical data, pharmacy data, patient-reported, behavioural, administrative and payer data, systems-level data) and be collected both prospectively and retrospectively. For example (and as discussed further in section 3), real-world data from electronic medical records, observational studies with prospective data collection and pragmatic trials is becoming increasingly relevant for medical research (supplementing data from RCTs) and enriching the evidence base on drug performance over time, as well as beginning to inform healthcare decision-making(Miani et al. 2014). Social network or ‘crowdsourcing’ data from patients who want to access and contribute to health information to support diagnosis, self-management and treatment monitoring, as well as to potentially inform the planning and provision of healthcare for a community, is also considered real-world data (Luciano 2013; Griffiths et al. 2012). As we elaborate on in section 3, different types of data present different types of value and will be used to varying degrees across research and innovation, care delivery and health systems or policy spaces.

Although both big data and real-world data are to a degree ‘umbrella’ terms, it is worth reflecting on their commonalities and differences. Big data describes the storage and analysis of large and/or complex datasets, and as with real-world data, often inheres in relatively unstructured data sets for source information. While there is overlap between the terms, the focus of the definition of real-world data is on the methods through which it is created (i.e. outside the framework of an RCT) but that of big data is more operational, focusing on the characteristics of the dataset and the analysis that it makes possible (e.g. using software such as NoSQL, MapReduce and techniques such as machine learning). Of course, data from several real-world data sources when combined constitutes big data, requiring support from specific software and machine learning to discern patterns (Miani et al. 2014). However, some real-world data analytics can cover relatively small and structured samples that do not necessitate big-data-powered algorithms for efficient analysis.

Finally, health data can also be classified according to the stakeholder that generates it. Considering health data through this lens we identify the following core categories: (i) citizen-/patient-generated data using various digital devices or media, such as data from mobile devices, wearable sensors, social media, blogs,
remote health monitoring devices; (ii) healthcare professionals/provider data such as EHR, hospital performance data, admissions data; (iii) payer data on claims and costs; (iv) researcher-provided data such as from scientific studies conducted in academia or research institutes; (v) government-provided data such as epidemiological population and public health data; and (vi) industry-/private-sector-provided data – such as pharmaceutical industry data from clinical trials or pharmacovigilance data, pharmacy data on medicines sales or other private-sector (non-pharma) data on public behaviour and sentiment analysis.

3. Different uses of health data and associated value

The literature on health data generation and use identifies a myriad of potential social and economic benefits, spanning three broad categories:

- Benefits for various stages of the R&D process
- Benefits for pharmacovigilance and public health
- Benefits for healthcare delivery and the health ecosystem more widely.

We discuss each in turn. As will be evident from the below the literature focuses more on future than realised potential, although some examples of already-realised benefits from enhanced data are presented as well. Appendix C provides additional detail on examples of health data initiatives.

3.1. Potential benefits for the R&D process

Several potential benefits from the use of health data in R&D (early R&D as well as later-stage trials) have been discussed in the literature. These include impacts on the questions which can be examined (research feasibility), operational and cost-efficiencies and research quality. Table 2 summarises the key types of value which could accrue and these are discussed in more detail in the narrative that follows.
### Table 2. Potential value of health data for research and development

<table>
<thead>
<tr>
<th>Opportunities to explore new research and innovation areas</th>
<th>Operational and cost-efficiencies in R&amp;D</th>
<th>Research quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Linking datasets and using new technologies to reveal patterns of association or causation that were previously not possible (e.g. understanding causes of Amyotrophic Lateral Sclerosis (ALS) in Finland by linking high spatio-temporal resolution data with epidemiological data)</td>
<td>• Better targeting of R&amp;D investments due to improved patient/population stratification, enabled by data on epidemiological profiles, genetic traits and clinical records (e.g. efforts of Genomics England Genomic Medicine Centres [GMCs] and Genomics England Clinical Interpretation Partnerships [GeCIPs])</td>
<td>• Greater external validity of research and enhanced confidence in results enabled by real-world data and by pragmatic trials that are more representative of real-world practice (e.g. Salford Lung study on chronic obstructive pulmonary disease [COPD] treatment)</td>
</tr>
<tr>
<td>• Enabling new types of health services research (e.g. Achmea Health Database, a registry of healthcare procedures in the Netherlands with prospects to inform quality-of-care evaluations and health services research)</td>
<td>• Operational efficiencies in clinical trials due to more precise recruitment, retention (hence reduced attrition) and site-selection strategies due to better patient-level data (e.g. the Electronic Health Records for Clinical Research [EHR4CR] initiative)</td>
<td>• Enhanced research quality due to prospects for new outcome measures over time (e.g. based on adherence and compliance data) which can inform future R&amp;D process</td>
</tr>
<tr>
<td>• Linking databases to enable greater sample size and statistical power in clinical trials for rare disease R&amp;D (e.g. 100,000 Genomes project in the UK)</td>
<td>• Reduced duplication and wastage of research due to access to richer and broader evidence base and enhanced data sharing. This also facilitates reproducibility and greater transparency and is in line with efforts to enhance public confidence in findings (e.g. GlaxoSmithKline [GSK] Clinical Study Register)</td>
<td>• Enhanced and more meaningful engagement of industry with patients in clinical trial processes, facilitating more evidence-based research responsive to patient needs and respectful of patient expertise and patient-reported data (e.g. PatientsLikeMe® portal-enabled engagement)</td>
</tr>
<tr>
<td>• Enabling research not possible under RCT model due to ethical considerations, but possible through real-world evidence and pragmatic trials</td>
<td>• A shorter timeline to regulatory review, marketing authorisation and commercialisation due to more efficient research management, enabled by more targeted, adaptive and evidence-based R&amp;D processes (e.g. European Medicine Agency’s [EMA] adaptive pathways pilots such as for a gene therapy for inherited blood disorder)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Contributions to volume and speed of generating research knowledge, for example due to a critical mass of clinical data (e.g. General Practice Research Database [GPRD] and Clinical Practice Research Datalink [CPRD] in the UK enabled a critical mass of publications and have informed National Institute for Health and Care Excellence [NICE] guidance on cancer and digestive disease)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Efficiencies brought by technological advances (e.g. genome sequencing)</td>
<td></td>
</tr>
</tbody>
</table>
3.1.1. Opening prospects for new types of research and new evidence through access to richer data sources and novel analysis techniques

Several publications highlight that linking datasets or making use of new technologies and analysis techniques can enable research that could reveal associations and causation that were previously difficult to detect (Corsham Institute and RAND Europe 2016; Couillard et al. 2015; INT01–12; Miani et al. 2014). To this effect, Sabel et al. (2009) discuss the example of research conducted in Finland that used high spatio-temporal-resolution data available for individuals, and through longitudinal research explored potential environmental or genetic explanations of ALS. This research uncovered implications for disease causation due to place of birth which would otherwise have been very hard to detect and could have remained hidden (Sabel et al. 2009).

Another example of using health data to create new research prospects (e.g. for health services research and evaluation) is the Achmea Health Database, a registry held by the health insurance company AGIS. It contains all healthcare procedures carried out in the last ten years using the records of 1.2 million health-insured individuals in the Netherlands. Using this information could create opportunities for data mining to inform the evidence base on healthcare consumption patterns as well as evaluations of quality of care and service planning, provided regulation is in place to ensure appropriate use. For example, a 2015 project used AGIS standardised insurance claims to examine the impacts of a health and social care support intervention on use of General Practitioner (GP) care and hospital care in the Netherlands, and identified recommendations and implications for a comprehensive integrated care approach to healthcare delivery (Kringos et al. 2016).

Using Magnetic Resonance (MR) imaging data for new analysis purposes is also gaining increasing attention in the health sector. For example, AMRA, a digital health company, transforms MR images into precise body composition measurements (e.g. showing fat tissue around body organs) by using a cloud-based, computer-aided service (AMRA 2017). This information, when used with other types of data such as Body Mass Index, can provide a fuller health status profile for individuals and be used in metabolic research and its therapeutic applications to predict disease risk or to provide more targeted methods of selecting clinical trial participants (AMRA 2017).

Access to new types of data also creates prospects for developing medicines for underserved patient populations for which medical solutions were not previously possible due to limited sample size. Linking different databases helps overcome critical mass and sample size issues, which increases the statistical power of clinical trials and reduces signal-to-noise issues (Costa 2014). This is particularly important for rare diseases or rare drug exposures (Auffray et al. 2016; EC 2014a; INT01). For example, the 100,000 Genomes Project (2012–present) in the UK aims to gather data from approximately 70,000 people. These are National Health Service (NHS) patients with cancer or a rare disease (and their families). Through a state-owned company, the project uses a secure, monitored infrastructure to store and analyse the collected data (genome sequences and clinical data).
3.1.2. More efficient and cost-effective R&D through improved, data-enabled targeting of investments and more stratified approaches

Literature on health data also discusses the potential of big data analytics and genomic research to enable more targeted R&D and hence more efficient processes of new drug target identification (Chataway et al. 2012; Couillard et al. 2015; Costa 2014; Morgan Jones et al. 2014). This could benefit both academic and industrial R&D efforts through more effective use of research funding. For example, ‘omic’ data (e.g. genomic, proteomic) and EHR data could enable researchers and industry to make better strategic decisions about future R&D development avenues due to a more precise stratification of the population based on needs and traits (Alyass et al. 2015; Costa et al. 2014). For example, the Structural Genomics Consortium (SGC), a partnership between public, private and third-sector actors, is using biomedical data in an innovative way and aiming to cost-effectively produce 3D protein structures on a large scale (Chataway et al. 2012). Through an open-access scheme, the Consortium is placing all the novel protein structures in a Protein Data Bank and also sending samples to researchers on request, for scientific purposes. In its first ten years of existence, the SGC discovered and published approximately 40 chemical compounds and produced 452 peer-review journals and eight books (Morgan Jones et al. 2014). The SGC model is also thought to lead to cost-savings. Within 2010, Biohub estimated that the SGC’s research into antibody effectiveness led to over $750m in savings (Morgan Jones et al. 2014).

An alternative example is reflected in the efforts of Genomics England and the associated GMCs and GeCIPs. GMCs aim to help enrol participants into Genomics England, accept NHS referrals and ensure appropriate sample and phenotypic information collection. As part of this initiative, genomic data will be related to phenotypic research data and clinical records to form a research dataset – for example to inform drug development, as well as to provide feedback to individual patients via their clinicians. The GeCIPs are being established as pre-competitive research networks that will take forward both specific and broad research agendas aimed at generating research to enhance drug development, and benefiting patients via improved diagnosis and treatment effectiveness (Dubow & Marjanovic 2015; Genomics England 2017).

The potential of health data to lead to benefits in later-stage medicine development and clinical testing has also been discussed in several publications. Some potential benefits include achieving operational efficiencies in running clinical trials by having more precise recruitment, retention and site-selection strategies through predictive modelling drawing on a variety of big data (Auffray et al. 2016; Couillard et al. 2015; EC 2016; Hood & Auffray 2013; INT01; Kalra et al. 2016; Mello et al. 2013; Salas-Vega et al. 2015). For example, Kalra et al. (2016) highlight the potential of EHR specifically to help with research design and implementation. They discuss how the EHR4CR initiative developed an innovative technological platform to securely connect to data in various hospital EHR systems and clinical data storage warehouses in Europe (in compliance with EU legislation). The platform aimed to provide clinical-trial sponsors with predictive analytics on the number of eligible patients for particular trials, to help identify appropriate trial sites, and to create a channel for connecting trial sponsors with healthcare providers able to assist with the recruitment process.

Analysts and scholars have also considered the potential of health data to influence the pace of R&D, and to potentially shorten timelines to market (Auffray et al. 2016; Geissbuhler et al. 2013; Morgan Jones et al. 2014). In terms of research speed, the GPRD in the UK linked data from primary care and enabled more rapid and more comprehensive research, yielding over 850 peer-reviewed related publications (as of 2011, Williams et al.
In 2012, the GPRD was incorporated into the CPRD. The CPRD is a joint venture between the Medicines and Healthcare products Regulatory Agency (MHRA) and the NIHR (National Institute for Health Research) in the UK. It aims to facilitate both observational and interventional research by offering access to anonymised patient data. Since 2012, over 640 research papers using CPRD data have been published and findings from CPRD-enabled research have informed NICE guidance in areas such as cancer and digestive disease (CPRD 2017; Kousoulis et al. 2015). By making use of patient data in a reliable, standardised way, the CPRD has also contributed to improving the research infrastructure and enabled studies that are expected to further improve the health and care system (Morgan Jones et al. 2016). Literature also highlights the decrease in costs of research enabled by technological advances such as genome sequencing (e.g. Costa 2014; Dubow & Marjanovic 2015), which are likely to bring further efficiencies to stratified medicine R&D in particular.

Another example of the potential of health data to impact on R&D efficiency can be found in the use of real-world data in more dynamic medicine approval schemes, as demonstrated by the EMA adaptive pathways pilots (e.g. for LentiGlobin BB305, a gene therapy for dependent beta-thalassaemia, an inherited blood disorder). This also creates prospects for faster regulatory approvals and shorter R&D timelines to market entry (EMA 2016). The adaptive pathways scheme allows for early and progressive access to a medicine to a limited sample of patients through conditional approval, with staged approvals and enhanced access prospects depending on emergent evidence on the benefit-risk balance of a product. In this approach, real-life evidence complements data from clinical trials (EMA 2017).

Data-driven operational efficiencies in R&D are also expected to accrue through greater sharing of research and wider access to data. This might not only minimise unnecessary duplication of research (e.g. Morgan Jones et al. 2014) but also enhance the evidence base for evaluating medicines and healthcare interventions. These benefits would be driven by increased accuracy, completeness and validity of the evidence (Koenig et al. 2015) and by improved public trust in clinical research and medicines due to enhanced transparency and ability to replicate findings (Loder & Groves 2015). One example of such data sharing efforts is the GSK Clinical Study Register, a data repository with protocol summaries, scientific results summaries, clinical study reports and complementary information from GSK-sponsored clinical trials (GlaxoSmithKline 2017). The BMJ Open Data Campaign, which requires publication in the BMJ to be predicated by access to anonymised patient-level data underpinning trials of drugs or devices, is another example (BMJ 2017).

### 3.1.3. Enhancing research quality

Literature also highlights the prospects of real-world data to enrich the quality of research evidence and especially external validity, confidence in results and targeting of findings to specific populations (Miani et al. 2014). Real-world data can also enable research which would not be feasible under an RCT model (Makady & Goettsch 2015) and provide new outcome measures (Raghupathi & Raghupathi 2014). According to one interviewee, it may also be easier to engage patients and the public with research through a real-world data model that offers a broader set of services around their patient journey (INT05).

Annemans et al. (2016) and Miani et al. (2014) identify that the potential for real-world data to inform clinical development is particularly important due to real-world data reflecting the outcomes for people in real-life situations rather than in a clinical-trial environment. Makady and Goettsch (2015) highlight that real-world data holds prospects for enabling greater external validity (generalizability of research findings). They also argue that Real-World Studies (i.e. studies which use real-world data – RWS) or pragmatic trials have more inclusive criteria for trial subjects and are hence more representative and more able to capture the heterogeneity of
treatment effects in clinical practice. They also span a wider time horizon than RCTs, and their results can be inferred to future effects with potentially more confidence than RCT results. According to Makady and Goettsch (2015), RWS could also facilitate the identification of subpopulations with better benefit–risk profiles, inform the design of pivotal trials, identify new disease relationships and inform new therapeutic targets. Garrison et al. (2007) discuss the prospects of real-world data being used in situations where RCTs are not ethical (such as conducting research on narcotic abuse) or feasible. In addition RWS could produce additional outcome measures like adherence and compliance parameters in a real-life setting, as well as reveal longer-term clinical events (INT08; Raghupathi & Raghupathi 2014). For example, the Salford Lung Study used real-world outcomes data from patients with COPD to identify new evidence of effectiveness of treatments in real-world settings for reducing COPD-associated exacerbations. This study was conducted through collaboration between industry, healthcare providers, community pharmacists and patients in Salford and the surrounding Greater Manchester area. It enabled the inclusion of a broad patient population in a pragmatic, digitally enhanced clinical trial design (University of Manchester 2016). However, capturing the benefits of real-world data will also require considering their feasibility and scale-up potential, taking into account the associated timelines and resource demands.

Data from EHR (directly from individual patients, or meta-analyses from social media data and patient platforms such as PatientsLikeMe and others) could also help facilitate more effective and meaningful industry engagement with patients who participate in trials. This data could enable a better understanding of patient experiences, symptoms and health behaviours (INT01). It could also support better-informed recruitment and participant retention in trials, as well as raise awareness about research needs and opportunities (INT01). One example of using health data in this way can be found in the online PatientsLikeMe Epilepsy Community Platform. The epilepsy platform within the wider PatientsLikeMe portal was established in January 2010 in the United States and in partnership with UCB Pharma, including its European branches. This platform allowed patients with epilepsy to record, monitor and share their data for the purposes of this research that aimed to increase the knowledge of patient perceptions and understanding of epilepsy. Data consisted of demographic, disease and treatment characteristics of patients, which enabled a greater understanding of epilepsy patients’ perceptions about the effects of their treatments and their quality of life (de la Loge et al. 2016). Based on retrospective analysis, the study revealed that patients reported a high rate of memory problems, fatigue and problems concentrating, as well as poor quality of life associated with these symptoms, side-effects as well as tonic-clonic seizures. The research led to a recommendation for a more holistic treatment approach beyond seizure control. The study also showed the research potential of the platform in providing patient-reported data (de la Loge et al. 2016).

### 3.2. Benefits for pharmacovigilance and public health

The use of health data for pharmacovigilance, disease prevention, health promotion and emergency preparedness is also discussed in both scholarly and grey literature (see Table 3 for a summary).
Table 3. Potential value of health data for pharmacovigilance and public health

<table>
<thead>
<tr>
<th>Category of benefit</th>
<th>Nature of envisaged benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved pharmacovigilance</td>
<td>• Linking real-world datasets (such as data from the primary healthcare setting and other specialised disease registries) provides more complete safety profiles for drugs over time and more realistic assessments of possible adverse events due to scale of data, thus increasing patient safety. <em>Examples of use: GPRD and CPRD, Exploring Adverse Drug Reactions by Integrative Mining of Clinical Records and Biomedical Knowledge (EU-ADR), Safety of non-steroidal anti-inflammatory drugs (SOS) project and effects on cardiovascular and gastro-intestinal health; Vaccine Adverse Event Surveillance and Communication (VAESCO) project.</em></td>
</tr>
</tbody>
</table>
| Enhanced, data-enabled public health, disease prevention and health promotion | • Larger and integrated environmental, genetic and socioeconomic datasets to enable better prediction of risk factors (*e.g.* NHS Atlas of Healthcare Variation).  
• Data on health apps and portable devices can enable citizen empowerment and facilitates citizen engagement in maintaining good health. (This value could potentially be scaled up further in the future if data reliability, user-friendliness, interoperability/standardisation, privacy and security challenges are addressed). *Example: In Hungary, a national health app initiative is trying to harness prospects for better use of data connected through sensors and devices to inform national health promotion strategies (lifestyle and health system organisation).*  
• Computer algorithms to assist in screening of patients for diseases could enhance early diagnosis and prevent progression of disease (*e.g.* this has been applied in areas such as breast, lung and colon cancer among others, often as a complement to radiologist-based diagnosis). |
| Enhanced emergency-preparedness                          | • Better monitoring of disease outbreaks due to timelier symptoms data (*e.g.* from both health professionals and social media) and data matching disease outbreaks with covariates (such as environmental data from satellite sensors) could facilitate early warning systems and more proactive response and disease management strategies.  
• More direct active citizen engagement in the collection and communication of disease outbreak data (*e.g.* via social media and portable devices to central databases) to health authorities may be enabled in the future (with checks on reliability needed). |

### 3.2.1. Improvements in pharmacovigilance

Pharmacovigilance is concerned with monitoring the effects of medicines after they have obtained a marketing authorisation so that patient safety can be ensured. Data is already widely used in pharmacovigilance (and indeed drug safety monitoring would be impossible without data). In this sense, real-world data has been informing pharmacovigilance activity for a long time already. However, literature points to the prospects of scaling up the use of real-world health data inputs into pharmacovigilance activities (*Couillard et al. 2015; EC 2014a; EC 2016; Mathew & Pillai 2015; Mello et al. 2013; Salas-Vega et al. 2015; Trifiro 2014*). Access to more diverse and greater amounts of real-world data, coupled with more granular information on patient profiles, could facilitate quicker and more rigorous learning about how drug safety relates to particular patient groups over time.

Several initiatives exemplify this application and value from health data, for example:

- GPRD data (introduced earlier in this report; see section 3.1.2) has been used in real-world harm–benefit assessment and pharmacovigilance, pharmacogenetic studies and pragmatic RCTs. This type of research has been possible due to GPRD linkage with databases containing disease-specific registry data, including socio-economic data and death registration data. Pharmacovigilance-focused research
using GPRD data has found that thiazolidinedione antidiabetic agents rosiglitazone and pioglitazone are associated with an increased risk of hip, spine, arm, foot, wrist or hand fractures (Douglas et al. 2009).

- The EU-ADR project aimed to detect and analyse drug safety signals by using data from eight population-based administrative and GP databases from Denmark, Italy, the Netherlands and the UK (Consortia-pedia 2016; Trifiro 2014). EU-ADR used special software – Jerboa – that shared only aggregated anonymised data using a common data model, and validation exercises showed important benefits in identifying safety concerns (Consortia-pedia 2016). For example, inferring from the drug utilisation patterns observed within the databases, the EU-ADR network would have been able to – via simulation – detect an association between rofecoxib and acute myocardial infarction within two years of the marketing of this medicine. However, simulations also showed that even if the EU-ADR network expanded to ten times its current size, there would still be a need for improvement in the sensitivity of pattern detection (Trifiro 2014). This EU FP7-funded project ran from 2008 until 2012 and insights from it have been subsequently applied in other EC projects such as SOS and VAESCO (Gini et al. 2016).

- Another example is the ADR-PRISM project, which aimed to leverage data from patient discussion forums for the use of pharmacovigilance professionals. The project was financed by the French Single Interministerial Fund. Lessons from this project could lead to new hypotheses on patient-reported side-effects which currently are unknown or poorly documented (ADR-PRISM 2017).

- The EU FP7 SOS project investigated drug safety for non-steroidal anti-inflammatory drugs (NSAIDs), in particular the association between these and gastro-intestinal and cardiovascular effects of traditional NSAIDs and COX-II selective NSAIDSs (Woerdeman 2013). Meta-analysis had shown that clinical trials were not large enough to show convincing safety estimations for gastro-intestinal and cardiovascular effects, particularly in children, while meta-analysis from observational studies did not provide estimates for heart failure. Linking data from 8.5 million new NSAID users, the SOS database study was able to inform a decision model for choosing the least toxic NSAID in respect to gastro-intestinal and cardiovascular effects (Woerdeman 2013).

- The VAESCO project involved eight European countries and investigated a possible association between the use of influenza A(H1N1)pdm09 vaccines and narcolepsy. The results from the project have informed EMA decisions on the use of this vaccine and have provided important lessons to European Centre for Disease Prevention and Control and public health and regulatory agencies in the eight participating countries on safety monitoring and building a vaccine safety assessment network in the EU (Destefano & Velozzi 2012).

- The EU-ADR workflow is now informing the European Medical Information Framework (EMIF) project (Gini et al. 2016). EMIF is running from 2013 until the end of 2017 and is aiming to develop a common information framework for linking patient-level data focusing on two main research issues: (i) understanding mechanisms that predispose people to develop dementias (such as Alzheimer’s); and (ii) understanding the individual factors for obese people that can lead to the development of complications such as diabetes (EMIF 2017).
3.2.2. Potential for other public health benefits: disease prevention and health promotion

Outside of pharmacovigilance, various other potential benefits from using health data for public health are discussed in the literature. These include prevention, emergency preparedness, and health promotion within healthcare and non-healthcare settings.

For example, research based on rich datasets – real-world data from patient registries, EHR or linkages between various types of datasets – enables more comprehensive identification of environmental, genetic and socioeconomic risk factors (Auffray et al. 2016; Wang & Krishnan 2014). Health data’s prevention value is also enhanced through the usage of computer algorithms to screen patients for diseases and enhance the accuracy of diagnosis (Costa 2014; Nguyen et al. 2012). Timely and accurate diagnosis could in turn help prevent the onset or progression of disease through early intervention.

Health information on variations in different health parameters (e.g. blood pressure, weight, sleeping patterns) is already beginning to contribute to empowering citizens and patients to engage in maintaining their health status, and may be able to impact on health promotion at a greater scale in the future (Mathew & Pillai 2015; Ping et al. 2014). This could be achieved for example through data communicated to patients, the public or clinicians via various health apps – provided that reliability is ensured and security and data privacy issues managed. Future prospects in this space may require a consideration of somewhat more standardised sensors and ways of feeding back information between healthcare providers and the public (via portable devices such as Apple Watch, iPhone or Fitbits), to ensure timely and evidence-based, actionable insights (Chen et al. 2012). These technologies gather data passively, which makes them convenient for the user. Although challenges to data reliability, data security and health professionals’ time and capacity to engage with such remotely connected data remain to be overcome to enable wider systemic level benefits (Chen et al. 2012), aggregating measurements from such devices and analysing them at population level could help inform wider public health promotion campaigns.

A recent example of putting these aims into practice is Hungary’s new national health app initiative (Collis 2016). The aim of the project is to improve public health by enabling good management of chronic diseases through lifestyle and treatment advice. The initiative is rooted in the concept of an app which connects users (patients/the public) to a central database of health data that can be shared with doctors if the user wishes to do so. The free app already has 20,000 users, who can upload information on their diet, weight, blood pressure, glucose levels, exercise and conditions such as pregnancy. A user in turn receives tailored advice to better manage their health. Although the information is centrally collected and stored, this initiative adopts an ‘opt out of data sharing’ policy meaning that patients/users can choose which healthcare workers to share their data with. According to Collins (2016), signs of impact are already beginning to emerge from the government’s efforts to make use of data through a centralised approach (extending beyond the example of just the app).

For example, through observing a trend in the treatment pathways for cancer patients, the Hungarian authorities were able to detect inefficiencies in charging and reimbursement for cancer therapy and to adapt cancer patient pathways. The Hungarian National College of Surgeons communicated that in 18 months, the number of liver metastases operations increased by 50 per cent as a result of better detection of patients requiring this surgery (Collins 2016).

Another example of the potential application of data for public health is exemplified by the NHS Atlas of Healthcare Variation, which is demonstrating the potential application of greater amounts of and more diverse data in public health.
This is a project using interactive mapping software to better understand and detect geographical determinants of health such as deprivation, lifestyle and environmental factors (Exeter 2014). Having produced already three compendiums (editions) with specific indicators of variation, the initiative is a useful tool to formulate key questions about equity, effectiveness and value in healthcare provision which could later inform healthcare delivery and public health campaigns in the prevention and health promotion space (NHS RightCare 2017).

3.2.3. Emergency preparedness

In the field of emergency preparedness, Salasl Vega et al. (2015) and Hoffman and Podgurski (2013) discuss that a wider usage of timelier health data (e.g. data from aggregated individual data, EHR, social media data) could lead to more robust monitoring of infectious diseases and disease outbreaks and facilitate early warning systems and more proactive response and disease management strategies. A 2014 report from the EC highlights that current disease mapping models could be improved by matching disease outbreaks with covariates such as environmental information gathered through satellite sensors (EC 2014; Hay et al. 2013). Furthermore, in the midst of a public health emergency or disaster situation when reaching physicians’ offices or computers is problematic, interoperable EHR systems could enable access to medical information on victims (Hoffman & Podgurski 2013). An expert we spoke to also highlighted that although big data is already used to monitor some outbreaks (e.g. flu), wider scale and more direct active citizen engagement in its collection and communication to healthcare professionals (e.g. via apps on mobile phones) could facilitate timelier decision making (INT01). Through this type of data sharing and interconnectivity, administrative efficiencies at disaster scenes could also be improved.

3.3. Benefits for healthcare delivery and the health ecosystem more widely

The literature also presents several potential benefits of health data for healthcare delivery and the wider health system (see Table 4). These benefits span potential impacts on the quality of care that can be delivered, operational and cost-efficiencies in care delivery, and wider benefits for health policy, health systems strategy and systems planning, drug reimbursement and the commissioning of care. Many of these benefits are currently realised at a relatively small scale as part of pilots or national initiatives, with prospects for wider-scale value generation in the future, as health data ecosystems mature.
Table 4. Potential benefits for healthcare delivery and the health ecosystem more widely

<table>
<thead>
<tr>
<th>Category of benefit</th>
<th>Nature of envisaged benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benefits for healthcare quality</td>
<td>• Better diagnostic capabilities which would draw on big data that correlates diseases and diverse biomarkers and symptoms with different patient profiles; improved screening algorithms; and integration of imaging data and symptoms data from EHR • Better and more personalised clinical decision-making (more effective outcomes, timelier treatment, fewer errors), for example due to enhanced predictive analytics capacity and more stratified approaches to treatment decisions enabled by more comprehensive clinical datasets <em>(e.g. EuResist Initiative)</em> • Workforce capability-building through care providers’ access to improved and richer evidence and new research knowledge on diseases and treatment options (provided that technological capacities and dissemination infrastructure ensure user-friendly dissemination)</td>
</tr>
<tr>
<td>Operational and cost-efficiencies in care delivery</td>
<td>• Greater transparency and cost-efficiency gains in the health system could result from easier comparability of outcomes from different treatment decisions across patient profiles, enabled by large clinical datasets such as EHR. Cost-efficiencies would also stem from reducing wastage and costs associated with administration of inappropriate or inferior treatment regimes • Health system efficiency gains (such as a reduction in unnecessary hospitalisations) through data-enabled self-care and self-management of risk factors and through remote monitoring of adherence to treatments could be partially enabled by ‘smart’ technologies, health apps, portable devices and empowered citizens <em>(e.g. ParkinsonNet self-care initiative in the Netherlands; BioSerenity and Dataiku’s Neuronaute)</em> • Faster information exchange between patients and physicians and between health authorities and clinicians – enabled by internet and mobile technologies and portable devices – could contribute to more accurate and timelier treatment decisions based on better information <em>(e.g. Finland stroke response model; IBM’s Shared Care Platform; Cloud computing at London’s Chelsea and Westminster Hospital; Optimum Patient Care Research Database [OPCRD]</em>)</td>
</tr>
<tr>
<td>Other envisaged benefits for the health ecosystem more widely; health service planning, strategy, regulation, reimbursement and commissioning</td>
<td>• Real-world pharmacovigilance data and outcomes data (clinical, patient experience) could enable better-informed drug safety regulation, adaptive pathways and reimbursement models. Greater usage of EHR could ultimately reduce cost at a systems level by facilitating more efficient health systems planning and improved workflows, productivity and administration efficiency. More aggregated data on different populations (e.g. demographic characteristics, drugs prescriptions reimbursed by the national health system, discharge records, outpatient visits and diagnostic-therapeutic assessments, pathologies) and cost data could for example be used for better-targeted resource allocation and planning of services <em>(e.g. cost data enabled by linked Italian regional databases identified the resource needs for managing acute coronary events)</em></td>
</tr>
</tbody>
</table>

3.3.1. Benefits for healthcare delivery and the quality of care

The literature points to a diversity of benefits from using health data for enhancing care quality through better clinical decision-making and workforce capabilities.

Belle et al. (2015), Couillard et al. (2015), Exeter (2014) and Mirnezami et al. (2012) consider prospects for better diagnostic capabilities which would draw on big data from a diversity of databases, such as data correlating diseases and diverse symptoms with different patient profiles. For example, new developments
such as multidimensional signatures\(^9\) could enable higher predictive power (e.g. for disease progression or treatment outcomes) than afforded by current practice which only uses biomarkers for some diseases (Auffray et al. 2016). The integration of medical images with EHR data and genomic data could also facilitate improvements in accuracy and reduce the time taken for a diagnosis (Costa 2014).

Another key potential benefit for healthcare quality discussed in the literature is in *better-evidenced clinical decision-making which would facilitate higher-quality, timelier treatment and fewer medical errors* (Costa 2014; Couillard et al. 2015; EC 2016; El-Gayar & Timsina 2014; Murdoch & Detsky 2013). If clinical decisions were supported at scale by data analytics and decision-making algorithms (for example through using large patient datasets and predictive analytics), then capacity to *forecast individual patient outcomes* could be enhanced (Cohen et al. 2014). These predictions could vary from medication courses to rehospitalisation forecasts (Mathew & Pillai 2015) and could influence wider care pathway choices. Descriptive analytics could also help identify high-risk patients who would be targeted for specific healthcare interventions, enabling more precise and personalised care (Cohen et al. 2014).

One example of health data’s potential to improve the quality of care is the EuResist Initiative, a project that builds on collaboration between pharmaceutical companies, governmental institutions, private companies and other partners. The project’s new technologies and mathematical models aim to provide a more efficient way to select the best drugs and drug combinations for any given HIV genetic variant by predicting responses to treatment. EuResist aims to create an online system that helps doctors choose the HIV treatment with the highest probability of halting virus replication and the evolution of drug resistance (Miani et al. 2014).

Another example is the use of the OPCRDR, which contains information on patient-reported experience and anonymised longitudinal medical records data from over 550 UK general practices, as well as the results of patient-completed questionnaires (Optimum Patient Care 2017). Mining this data in research studies could facilitate efforts to develop new treatments and medical breakthroughs. For example, using patient-completed questionnaires from the OPCRDR, a research team that included researchers from GSK and Novartis investigated the relative efficacy of drugs used to treat COPD. The data and analysis led the team to conclude that COPD was not being treated according to the recommendations of the Global Initiative for COPD and the NICE, impacting the quality of care provided (Price et al. 2014).

Luciano et al. (2013) discuss that the increased applicability of health data could also bring about further *development of clinicians*. For example, more rapid, contextualised and widely disseminated knowledge could enable clinicians and health researchers to keep up to date with the most recent medical literature, provided that information could be filtered, processed, analysed and communicated in a user-friendly manner without imposing undue burden on health professionals’ time (Luciano et al. 2013). Wachter (2016) also highlights the importance of user-friendly design of health IT systems.

### 3.3.2. Potential efficiencies in healthcare delivery

Greater transparency and access to data could facilitate easier comparability of outcomes from different treatment decisions across distinct patient profiles (e.g. patients with different genotypic, phenotypic and behavioural traits), enabled by large clinical datasets such as EHR (Couillard et al. 2015; INT01). More effective

---

\(^9\) Clinical signatures are the patient’s disease representation at a certain moment, characterised by symptoms and quantitative or qualitative measures including expressions for a biomarker (Joyce et al. 2017). Multidimensional signatures take into account prior information on the patient’s medical history and from different medical databases (Auffray et al. 2016) as well as up-to-date information on the disease incorporating the dimensional definitions of disorders (Joyce et al. 2017).
treatment protocols based on richer data could also facilitate efficiencies by reducing wastage and costs associated with administration of inappropriate or inferior treatment regimens and care pathways (EC 2014a; Matheson J, 2015; Raghupathi & Raghupathi 2014).

Literature also identifies potential for health system efficiency and effectiveness gains through data-enabled self-care (Couillard et al. 2015; Raghupathi & Raghupathi 2014; Salas Vega et al. 2015). For example, by gaining greater knowledge about self-management and risk factors, patients may be in a position to make more informed choices about whether to remain in their homes (and avoid the expenses of hospital or nursing home-based care) in collaboration with health professionals (INT06; Salas-Vega et al. 2015). Of course, such approaches would require careful risk management by the health system to ensure that necessary hospital care was not omitted. Sensors/‘smart’ technologies and virtual communication platforms could facilitate self-care approaches, provided that regulatory and technological challenges to their use are addressed. Pattern detection through real-time wearable sensors for elderly or disabled patients could also alert the physician about high-risk situations that would require immediate prevention actions (Couillard et al. 2015). For example, BioSerenity and Dataiku’s Neuronaute wearable device aims to improve the diagnosis and real-time monitoring of epilepsy. Data gathered through this device could – in principle – provide doctors and care providers more access to information, thereby allowing quicker reaction rates and potentially improved care outcomes (News-Medical 2016).

Another example of health system support for self-care strategies can be found in the experience of the ParkinsonNet (2017) initiative. This initiative is facilitating high-quality multidisciplinary care for people with Parkinson’s disease by rapidly creating and disseminating new knowledge within the healthcare system. Through a web-based platform, ParkinsonNet facilitates interaction between professionals and patients, encouraging the transfer of information on best practices (Bloem & Munneke 2014). The system originated in California and is now implemented in all Dutch municipalities as well as in parts of Norway. An evaluation of the initiative in the Netherlands found it provided both significant cost savings (€20m annually, equating to four to five per cent of the total expenditure on direct Parkinson’s care in the Netherlands) and reduced the occurrence of hip fractures in patients with Parkinson’s disease by 55 per cent (Bloem & Munneke 2014; Nijkrake et al. 2010). In a wider study on potential benefits, PricewaterhouseCoopers (2013) estimated that by using mHealth solutions, healthcare systems in the EU could save up to €99bn in total annual healthcare spend in 2017 and help 185 million patients gain 158,000 years of life (PricewaterhouseCoopers 2013). However, this would assume overcoming a diversity of systemic and operational barriers to adoption.

Faster information exchange between patients and physicians and enhanced communication between health authorities and clinicians is also discussed as a source of efficiency and effectiveness gains in the provision of care. For example, dynamic interaction between various patient-controlled platforms as well as professional networks could lead to timelier, more accurate and more relevant information (Couillard et al. 2015; Griffiths et al. 2012). To illustrate with some examples:

- In Finland, health professionals have developed a model for stroke patients to receive quicker treatment, using mobile phones for urgent communication with health and care professionals as well as state-wide integrated EHR (Matheson R, 2015).
- Another example is IBM’s Shared Care platform, a project based in southern Denmark, which involves two hospitals, three municipalities and several medical practices, as well as IBM. The main goal is to facilitate real-time communication between patients, physicians, pharmacists and mental-health
professionals, so that all parties have timely and comprehensive insight on a patient’s status (News-Medical 2013). This platform uses large volumes of data from the clinical setting, analyses the data to generate insights for healthcare providers, and integrates health records from health, social services and other providers (News-Medical 2013). The programme initially focused on heart disease and intends to expand to include patients with type II diabetes and pulmonary conditions (News-Medical 2013).

- The Scottish Single-Shared Assessment System (SSA) also presents a platform for sharing patient-level data (provided consent is given) in an efficient way and between health and care providers. It is based on the collection of diverse types of data, including on medical history, home environment and current medication, and unifies previously separate systems in finance, management and delivery of services. Data can be shared among people who use the service and the agencies and the professionals who deliver the service (King et al. 2012).

- The Innovative Medicine Initiative’s Remote Assessment of Disease and Relapse in Central Nervous System Disorders (RADAR-CNS) project aims to develop new means of monitoring major depressive disorder, epilepsy and multiple sclerosis with the help of wearable devices and smartphone technology (Innovative Medicines Initiative, 2017). It involves collaboration between diverse stakeholders and sectors (academics, healthcare professionals, industry, patients and citizens, clinical service providers, payers and regulators across digital and life-science sectors (Wicks et al. 2016). This programme started in April 2016 and is expected to last until 2021. It is co-led by Janssen and King’s College London.

3.3.3. Other envisaged benefits for the health ecosystem more widely

Finally, literature also highlights wider potential benefits at the health systems level from the usage of health data including a more comprehensive evidence base or enhanced predictive capacity. These benefits span contributions to: health policy, health systems planning and strategy for resource allocation; drug payment and reimbursement mechanisms which consider real-world data and conditional approval as well as enhanced post-marketing safety and pharmacovigilance; and insurance provision and the commissioning of services (Couillard et al. 2015; Makady & Goettsch 2015; Miani et al. 2014).

Various studies point to the potential of greater usage of EHR and their interconnectivities to improve workflows, productivity and administration efficiency and to ultimately reduce costs for health systems by facilitating more efficient health systems planning (Belle et al. 2015; Couillard et al. 2015; El-Gayar & Timsina 2014; Miani et al. 2014; NHS European Office 2016; van Panhuis et al. 2014). For example, more aggregated data on different populations (e.g. grouped by pathology, age or location) as well as cost data could be used for resource allocation (Costa 2014; Roggeri et al. 2013). In the context of drug pricing decisions, Makady and Goettsch (2015) discuss that drug reimbursement models could be improved by complementing pharmacoeconomic modelling with real-world data on costs, resource use and utility values. This implies a wider-scale, richer, broader and longer-term outcome and impact data strategy which could potentially be valuable in informing outcome-based healthcare financing models (Makady & Goettsch 2015) and the commissioning both of services and of specific treatments (INT01).

Finally, a wider employment of health data such as EHR could help with fraud detection. For example, Mathew and Pillai (2015) emphasise that predictive models based on decision trees, neural networks and regression could be useful to predict and prevent fraud at the point of transactions (Venkata et al. 2012).
3.4. Who could benefit from health data?

The value of health data can also vary depending on the stakeholder. Reflecting on the diversity of benefits discussed above, Table 5 highlights key types of prospective value for specific stakeholders in the health sector ecosystem, spanning: citizens and patients; healthcare providers; academia and the research community; industry; payers, regulators and policymakers; and charities/third-sector organisations.

Table 5. How different stakeholders can benefit from a health data-rich society

<table>
<thead>
<tr>
<th>Stakeholder groups</th>
<th>Illustrations of key types of value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citizens and patients</td>
<td>• More personalised care through data-enabled more targeted identification of risk factors and treatment approaches. Biological (e.g. genetic and proteomic data), clinical, sociodemographic, pharmacovigilance, wellness and environmental data offers prospects for a full life-cycle approach to personalised care and personalised medicine spanning prevention, diagnosis, treatment, cure and health promotion spaces (EC 2014a; EC 2016; El-Gayar &amp; Timsina 2014; Hood &amp; Auffray 2013; INT01–10; Mirnezami et al. 2012)</td>
</tr>
<tr>
<td></td>
<td>• Patient empowerment to engage with care decision-making facilitated through improved access to data via technology and data-enabled platforms (e.g. apps, internet, remote monitoring) which facilitate more timely communications with healthcare providers on possible care pathways (Griffiths et al. 2012; Hollis et al. 2015; Monteith et al. 2015; Murdoch &amp; Detsky 2013). An interviewee (INT06) highlighted that this may also enhance citizen autonomy. This is relevant in the context of ageing populations that experience a decline in the degree of autonomy and self-reliance, which can in turn affect mental well-being.</td>
</tr>
<tr>
<td>Care providers/health professionals</td>
<td>• Better-quality and timelier clinical decision-making and ability to deliver higher-quality patient care due to access to improved data on patient symptoms and links to disease, enhanced screening tools and diagnostic accuracy, predictive analytics for treatment outcomes and more integrated EHR reducing time between tasks (Costa 2014; Couillard et al. 2015; EC 2016; El-Gayar &amp; Timsina 2014; INT01–10; Murdoch &amp; Detsky 2013)</td>
</tr>
<tr>
<td></td>
<td>• Prospects for enhanced self-management and remote care outside of the hospital freeing up capacity and reducing costs on publicly funded systems/organisations (e.g. via remote monitoring, digital-media-enabled enhanced communication of health information and advice to patients) (Couillard et al. 2015; PricewaterhouseCoopers 2013; Salas-Vega et al. 2015)</td>
</tr>
<tr>
<td></td>
<td>• Prospects for professional development through access to a richer array of medical information and cutting-edge research evidence (El-Gayar &amp; Timsina 2014)</td>
</tr>
<tr>
<td>Academia and the research community</td>
<td>• Prospects for new types of analysis and for addressing research questions which previously were not feasible due to improved data on prevalence, disease trends and patterns of association and evaluation of risk factors based on large population cohort studies, as some examples (Corsham Institute and RAND Europe 2016; Couillard et al. 2015; Douglas et al. 2009; EC 2016; Miani et al. 2014; Woerdeman 2013;)</td>
</tr>
</tbody>
</table>
|                     | • Possibilities to influence research agendas of other fields. As illustrated by an interviewee: ‘Academia could benefit from wider access to a wider variety of data and the possibility of research communities to influence research in related fields, for example in the development of the technology itself....[The research community may be seen as a passive consumer of advanced information technology in health, and, having access to health data, they can influence technological developments such as supercomputing developments. And they do.’ (INT06) Another example would be the influence of genomic medicine on developments in the fields of ethics, cloud computing and data storage as well as high-throughput screening (Dubow & Marjanovic 2015)
| Pharmaceutical and other industry | **• Creating efficiencies in the drug development process through better targeting of R&D investments** (Auffray et al. 2016; Couillard et al. 2015; EC 2016; Hood & Auffray 2013; Kalra et al. 2016; Mello et al. 2013; Salas-Vega et al. 2015)  
**• Establishing products with an improved efficacy and drug safety profile** due to better predictive analytics for stratified patients, real-world outcome monitoring data and enhanced pharmacovigilance data (Couillard et al. 2015; EC 2016; Miani et al. 2014; Woerdeman 2013)  
**• Using real-world data throughout the R&D and innovation cycle could also facilitate reimbursement** for products that have a proven enhanced efficacy in a real-life setting and inform value-based and outcome-based payment approaches, as well as adaptive pathways and policies (Accelerated Access Review 2016; INT01; INT03)  
**• Other industries outside of pharma are also benefiting from the wider use of health data** (e.g. ICT data analytics, wearable technologies, energy companies). In addition to the companies that specialise in producing wearable devices, new ICT companies (e.g. Appistry, Beijing Genome Institute, CLC Bio, Context Matters) are providing solutions for data analytics aspects for the eHealth ICT sector (Costa 2014). There could emerge a more specialised clinical research ICT sector focusing on clinical trial and clinical care software and hardware. Other companies such as Amazon Web Services, Cisco Healthcare Solutions or Dell Healthcare Solutions are expanding their portfolio and offering big data services (e.g. storing and analysing biomedical information). Companies from the energy field are also becoming involved with health data aspects and contributing ambient data |
|---|---|
| Payers, policymakers and regulators | **• Prospects for better health service planning, management of healthcare demand and service commissioning**, for example through linking claims and health service usage data (e.g. on medical procedures, hospital stays and administered drugs), provider performance data and individual patient-level data from EHR (Belle et al. 2015; Couillard et al. 2015; INT01; INT10; Miani 2014; van Panhuis et al. 2014). This could also lead to more effective risk-management programmes (INT01). Healthcare policy design and strategy could be improved by making better use of studies taking into account a greater number of determinants of health, which could also inform prevention/wellness programmes (INT01)  
**• Access to real-world outcomes data and pharmacovigilance data could facilitate more evidence-based and dynamic reimbursement decisions and health technology assessments of cost-effectiveness** (Makady & Goettsch 2015), accommodating also changes in value assessments and pricing over time as a critical mass of evidence and stratified data accrues |
| Third sector and charities | **• Charities and other third sector organisations** (e.g. civil society and voluntary organisations) **could benefit from more evidence-based advocacy efforts and enhanced patient and public participation in framing advocacy agendas** (e.g. via social media data) (INT06)  
**• Access to better data on research needs could influence charity-based research funding** (INT06) |
4. Enabling receptive health data ecosystems

At present, the landscape of health data generation, interpretation and use is still in a relatively early phase of evolution, with potential for much further development to create more enabling conditions. As the examples presented in section 3 point to, many of the benefits of health data are currently realised at a relatively small scale. Whereas health data-related initiatives are in a phase of expansion, many of them are yet to translate into practice. In general, applications of health data (especially big and real-world data) for R&D purposes and for pharmacovigilance seem to be somewhat more developed than applications in healthcare delivery and health systems planning. The overall landscape is characterised by promising but relatively fragmented and piecemeal efforts to capture value. Nevertheless, these provide fertile ground for shared learning that can inform future progress, and especially in the context of navigating and responding to diverse challenges, as they evolve. A supportive health data ecosystem is yet to be established in order to harness the full potential of health data and scale up existing benefits at pace. To achieve this requires further concentrated action in a range of areas – reflecting both opportunities and the inevitable challenges that accompany transformational initiatives of the scale presented by a health data-rich society and economy (see Box 1). The literature and experts who informed this review have suggested several areas to consider in efforts to create sustainable and effective health data ecosystems. These include:

1. Considering the value propositions presented by health data in the context of interdependencies and interactions between stakeholders and sectors.
2. Building on existing efforts: momentum to address data quality and technical considerations.
3. Making the most of recent data protection advancements and considering models of data sharing.
4. Workforce capacity-building: supporting professionals in making the most out of health data.
5. Learning from previous efforts to increase public awareness, acceptability and engagement with health data.

Box 1. Towards receptive health data ecosystems – a summary of key messages

1. Capturing value from health data will require collaboration and coordination between multiple stakeholders and sectors (e.g. health and social care) to ensure the effective management of interdependencies between them and the alignment of interests. Efforts to scale up current health data initiatives need to engage individuals and organisations with boundary-spanning and brokerage roles to tackle structural, technological, cultural and historical boundaries to collaboration and data sharing.
2. Data quality assurance, harmonised standards, compatible IT infrastructure, technologies which facilitate aggregation and sense-making, secure forms of data storage and transfer, and the governance of legitimate data use will be critical for receptive health data ecosystems. EU-wide efforts are under way to lay the framework and provide guidance for standardisation and interoperability requirements.
3. Recent advancements in GDPR are paving the way for a clearer governance framework for anonymised health data use in research and care delivery, with simplified consent requirements and rigorous technical and organisational safeguards. The boundaries of what is considered research use and actor eligibility require further clarification. There is also a need for further collaboration and coordination on issues of GDPR implementation at EU and national levels – including to better understand member state attitudes to this implementation. Ensuring an understanding and acceptability of the GDPR governance framework among the wider public through effective
4. In addition to anonymisation, various models of data sharing and privacy safeguards (e.g. differential privacy and block-chain technology, trusted curators of data, dynamic consent models) are being considered and trialled in the context of the health sector. Further research is needed on the system-wide resources and capacities needed to implement and manage such safeguards, and on how they relate to GDPR framework efforts to simplify the system. Also, regulatory capacity will be needed to ensure sufficient oversight of data-sharing practices and the validity of the data to facilitate both commercial confidence (e.g. in terms of clinical trial data) and patient privacy.

5. Efforts to ensure public acceptability and engagement with health data initiatives call for learning from prior efforts, and especially regarding effective engagement and communication campaigns. Patients and the public need to understand the opportunities, risks and safeguards related to health data use, who has access to which data and why, portability issues, ways of handling data, opt-out options, potential impacts on doctor–patient relationships, wider public-good benefits and legal aspects that accompany a health data-driven system.

6. New workforce skills and competencies to engage with, interpret, communicate and act on health data will be needed by the research and care-delivery communities, as well as by regulators, payers, policymakers and the wider public. This needs to be considered in educational curriculums and continual professional development programmes, and in citizen empowerment and outreach efforts. Workforce capacity-building is likely to require a focus on both clinical informatics skills and more general leadership, management and communication skills.

7. No single initiative can resolve all of the challenges and harness all the opportunities for value generation and impact that a supportive health data ecosystem could create. There is a need more for a critical mass of evidence and more learning-focused and thematic evaluations to inform future interventions and to identify transferable lessons for scalability and sustainability.

4.1. Considering the value propositions presented by health data in the context of interdependencies and interactions between stakeholders and sectors

As discussed in section 3, the current knowledge base discusses benefits for diverse stakeholder groups, and society more widely, which a health data-rich society could enable if the technical, legal, ethics-related, governance and data protection-related, and cultural challenges affecting public acceptability are addressed. However, there is less consideration of how the generation of value for individual stakeholders depends on the interactions between them, and on their collective collaboration and coordination. Many reports have tended to focus on value for specific groups (be it industry, academia, patients, health providers, policymakers, payers, regulators or the third sector; please see section 3.4) or different contributors to the value pathway such as data generators, data services, data business users and end consumers (European Political Strategy Centre 2017). This provides helpful learning which could be taken further to examine and understand interrelationships and interdependencies and to examine brokerage and boundary-spanning roles. For example, as expressed by one interviewee we spoke to: ‘The industry can be part of this debate, and can be very much a matchmaker, that they bring together those who do research with those that can provide technological solutions so can provide access to data. It can serve as an intermediary to stimulate innovation that involves several parties as this becomes a cross-disciplinary activity.’ (INT06) In principle, other stakeholders could have complementary boundary-spanning and brokerage roles – for example policymakers as brokers between healthcare providers and payers. The European Institute for Innovation Through Health Data (Kalra et al. 2016) is seeking to address some of these coordination and collaboration challenges. Building on the results of the EHR4CR and SemanticHealthNet projects, the institute aims to foster engagement of several stakeholders and establish a research network to tackle bottlenecks associated with the use of health data.
Among them are best practices in information governance, harmonised health information and user solutions for better-quality data and legitimate use of health data.

There is limited consideration in the literature of the interdependencies between the health sector and other sectors (such as social care, digital economy and employment for example). As emphasised by an interviewee for this review: ‘This is becoming a cross-sectoral type of activity so it’s important to highlight that what you can do with data and health is very much influenced by policies coming from areas that are not necessarily health.’ (INT06) A supportive health data ecosystem would need to consider these issues to facilitate scale-up and sustainability. For example, in a healthcare paradigm that puts the patient at the centre, health and social care systems would need to complement each other’s competencies and avoid inefficiencies in collaboration, including potentially through more integrated models (King et al. 2012). One interviewee stressed that ‘health and social care data need to be treated in the same way, in terms both of the constraints put on them and the sharing’ (INT12). In the context of data-driven sectors, this would imply a need to overcome structural, cultural and historical boundaries to collaborative working and data sharing, and incompatible IT systems and infrastructure (King et al. 2012).

In addition, the introduction of technical systems to enable effective data sharing would need to be accompanied by effective governance and management, systems interoperability, good training and staff development programmes, and clear guidelines and regulations. Illustrating the importance of regulatory guidelines and a clear classification of medical devices, some interviewees consulted during this study (INT03, INT06, INT12) highlighted the importance of such measures in the context of medical device regulation. At present, smartphone apps with an explicit medical purpose and application (e.g. diagnosis) are classified as medical devices. To avoid risk and ensure compliance with medical device regulation, some companies classify their apps as medical devices, even if the precise medical nature of their application is unclear or debatable (e.g. some mobile health apps which can be focused on well-being more generally). In contrast, literature also identifies cases where companies choose to classify their products as apps rather than medical devices and therefore elude some of the requirements of data sharing that have been introduced by specialised journals such as the BMJ (Loder & Groves 2015).

Value-chain thinking and value-chain mapping may be helpful in exploring these dynamics (INT07). Value chains are a useful tool to describe and analyse all the activities that firms and workers engage in to bring a product from concept to its end use and beyond (Duke University 2016). Through this process one can identify the roles and responsibilities of, as well as the relationships and interdependencies between, different actors and sectors – and how they can support or challenge each other in efforts to enable supportive health data ecosystems. According to some participants at the expert roundtable conducted as part of this review, health data ecosystems will need to be underpinned by a social contract between industry, policymakers, health providers, research community and citizens. Among other factors, this contract would consider the benefits and trade-offs that come with different approaches to health data governance and management, such as different degrees of centralisation (Wachter 2016). According to experts consulted at the roundtable for this review, the benefits that are likely to accompany a decentralised approach relate to greater levels of trust facilitated by quicker and more frequent interaction between stakeholders, partially facilitated by geographic proximity. However, such an approach could disadvantage some communities (e.g. rare disease-related) where it is challenging to achieve a critical mass of data and evidence on which one could base decisions. A centralised approach would involve institutions at the highest levels and would be beneficial for ensuring and promoting a common set of goals. However, roundtable participants also noted that a centralised approach is challenging to
execute and deliver in practice, and that a hybrid model (in which clear principles are set at national levels, but implementation managed regionally and locally) may be more feasible.

4.2. Building on existing efforts: momentum to address data quality and technical considerations

The possibility to exchange and link data in a secure and confidential manner requires technical capacities and capabilities, in addition to a conducive data protection and wider governance framework (elaborated on in section 4.3). Ensuring data quality, achieving requisite degrees of standardisation and interoperability, providing secure forms of data storage and developing data analytics capabilities such as algorithmic accountability are the main technical issues discussed in the literature (Alyass et al. 2015; Auffray et al. 2016; Costa 2014; EC 2014a; Hoffman & Podgurski 2013; Matheson J, 2015; Montheith 2015; Trifiro 2014; van Panhuis et al. 2014). The technical abilities to store, share, assess and reuse information will also require appropriate infrastructure to meet the stakeholder concerns about ownership and control (Keen et al. 2011).

The power of health data in both the clinical and the real-world setting comes from its aggregation and ability to be interpreted and made sense of in an actionable way. For research and innovation, healthcare delivery and health policy purposes alike, this calls for data formats amenable to aggregation (i.e. a degree of standardisation) and comparison, interoperable systems and consistency in quality assurance (King et al. 2012). At present, the different types of EHR used throughout the EU make this challenging (Auffray et al. 2016). In addition, efforts to ensure anonymisation or pseudo-anonymisation can be in tension with efforts to enable dataset linkage and interoperability, and can compromise the value and usability of the data.\textsuperscript{10} Probabilistic matching techniques such as Bloom Filters (Schnell et al., 2009) are trying to address challenges to data linkage while ensuring that findings do not result in data becoming identifiable. Mirnezami et al. (2012) highlight that achieving a critical mass of health data would also call for adaptations in already-established disease classifications, for example updating to take into account molecular information.

Standardisation and interoperability efforts are under way and aim to clarify the requirements, specifications, guidelines and characteristics that can be used to link and integrate different datasets. To illustrate with some examples:

- The World Health Organization (WHO) European Health Information Initiative (EHII) is bringing together partners to work on harmonising health indicators and developing an integrated Health Information System for the European Region (WHO 2017).
- Data mining techniques are being advanced and improved to extract information, and new analytical frameworks and methods are being developed (Belle et al. 2015; Cohen et al. 2014; Costa 2014; Holzinger & Jurisica 2014).
- Cloud computing systems offer enhanced prospects for storing data, if security and reliability can be assured (Sultan 2014).
- The EHR4C initiative (see section 3.1.2) – a recently completed project (2011–2016) that involved collaboration between 35 academic and pharmaceutical companies, as well as 11 hospital sites in France, Germany, Poland, Switzerland and the UK – established simulation techniques that could

\textsuperscript{10} This is partially because of challenges to ensuring that data interpretation using different sources is based on data on the same individual, which is sometimes needed.
analyse anonymised data and provide information to identify eligible patients for trials, assess a trial’s feasibility and locate the most relevant hospital sites (EHR4CR 2016).

- The ongoing EMIF project is building a catalogue of data that aims to provide researchers a single point of access for searching aggregated data across different sources and countries, from consented users (EMIF 2017).

- The TRANSFoRm project in 2010–2015 was sponsored by the EC FP7 programme, and used integrated EHR from Europe’s GPs with the aim of developing a core, integrated computational infrastructure. The TRANSFoRm project aimed to promote interoperability between different clinical systems across different countries, as well as aid in improving clinical decision-making (i-HD 2016).

- The EU Smart Open Services for European Patients (epSOS) Pilot Project aimed to create cross-border data exchange by building an ICT infrastructure to allow different European healthcare systems to access patient health information (epSOS 2017).

- Other EU-level proposals include developing a data community through a Public-Private Partnership (PPP) model that would for example foster best practices in sharing of datasets, develop incentives for knowledge and technology transfers, and launch incubators to accelerate digital entrepreneurship (EC 2014c).

- The EU GDPR (discussed in more detail in section 4.3) considers and introduces a requirement for safeguards in terms of technical and organisational measures for data minimisation (processing the minimal amount of personal data) (GDPR 2016).

- The usage of guidelines such as the Continua Guidelines, which reference standards and specifications to ensure interoperability of health IT solutions. This is currently employed in Denmark (INT12).

Real-time evaluation of these efforts could be informative for targeting future directions and acting on learning in a timely manner. No single initiative on its own can resolve all the challenges and harness all the opportunities, and it will be important to ensure knowledge-exchange, collaboration and cooperation between them.

4.3. Making the most of recent data protection advancements and considering models of data sharing

In April 2016, the European Parliament and Council adopted the GDPR, providing a legal framework for the processing and usage of personal data in the EU. The GDPR is set to come into force in May 2018. Within GDPR, health data is viewed as sensitive data which can be collected and used without special consent for the purposes of preventive or occupational medicine, medical diagnosis, provision of health or social care or treatment, management of health or social care systems and services — when these involve a health professional or person who would be bound by professional secrecy under law (the ‘medical care’ ground). Additional exemptions from consent are allowed when data is necessary for public health reasons (the ‘public health’ ground) or scientific research. However, there is no clear definition of scientific research in the GDPR (Hordern 2016), and this is an area which is likely to benefit from clarification of scope, actor eligibility and
boundaries to what is considered research as opposed to commercial activity, and what constitutes compatible and secondary data use.

Safeguards to ensure that technical and organisational measures are in place for data minimisation (processing the minimal amount of personal data), the right to erasure (i.e. the right to be forgotten) and the right to object to data processing (in particular circumstances) are also required by the directive. In regard to data ownership, Article 20 of the GDPR specifies the right to data portability for the data subject (e.g. patient/member of the public), who can transmit such data to another controller (e.g. a health professional).

The GDPR represents a long and rigorous effort of negotiation and agreement between diverse stakeholders at EU level and provides a comprehensive, unified and simplified EU data protection framework, while permitting a limited degree of country flexibility. For example, as illustrated by one expert we spoke to: ‘Some countries have flexibility on overrides and cases when consent does not have to be sought. So we have to be careful to make sure that people understand that it’s an override for the purposes of protecting people providing care to them. We need to facilitate an understanding among the public of what the laws are in different countries [not least] because there are people who move.’ (INT03). Currently, the data sharing space is still characterised by fragmentation and challenges to cross-border data linkages (Couillard et al. 2015; Geissbuhler et al. 2013; INT01–10; Stril et al. 2015). In the words of an interviewee: ‘So if you are trying to combine your health datasets across Europe…not necessarily allow you to transfer that data across territories – so that limits your ability to use these types of datasets and come forward with…innovative solutions to problems that are facing patients across the continent. There are some objective attempts to ease that problem with more pan-European legislation…for harmonising the way you have to deal with this data, but at the same time this legislation still allows member states to put their own specific rules or further exemptions on health data, how you go about processing it and reprocessing it and sharing it.’ (INT02)

Fragmentation could also exist at country level with states that have different regional requirements (INT08). The implementation of the GDPR requirements at country level will also require countries to develop internal capabilities and capacities when it comes to data controllers, and to ensure funding for sustaining such bodies (Matheson J, 2015). Furthermore, accountability mechanisms should be ensured throughout all member states; as expressed by one interviewee, ‘the Data Protection authority is relying on more of a self-reporting approach with Data Administrators only needing to provide in writing the procedures they will follow’ (INT11).

As expected with an initiative of this scope and scale, its implementation is yet to unfold and there are diverse areas for further consideration as the initiative evolves. In addition to clarifying what falls in the scope of data use for research, GDPR implementation will require a better understanding of: (i) how EU member states will react to the guidance issued; (ii) how protection of personal data will be ensured when data is transferred and processed outside the EU (EC 2017); and (iii) what regulatory capacity has to be in place to ensure sufficient oversight of data access and sharing practices. Insights from participants at the expert roundtable also suggest that an active campaign to ensure public understanding and acceptability of the GDPR governance framework will be important to facilitate engagement and contributions to value generation. Some participants noted that regulation and successful translation of GDPR in practice is likely to involve a mix of policy, technical regulation and standards, as well as self-regulation through community norms and behaviours.

In addition to institutional responses such as the GDPR, the literature offers several technical suggestions for addressing data sharing and access issues specifically. For example, Morley-Fletcher (2014) discusses a differential privacy model in which sensitive information about individuals is protected, not only through
anonymisation, but also through an algorithm and an interface that serves as a trusted curator. In this approach, the curator supplies the data to the researcher after having injected some randomisation into the answers, protecting the privacy of the individuals in the dataset. Morley-Fletcher (2014) flags that some patient coalitions or cooperatives are already emerging that entrust their data to a curator responsible for safely storing and managing all health data in individual accounts (Morley-Fletcher 2014). Subsets of this aggregated data are then shared with other actors, depending on the nature of different agreements. An example is Midata.coop, an initiative that enables citizens to store their information in one secure place. The citizens can also decide whether they wish to share this data with healthcare providers or to participate in research and grant access to subsets of their data (Midata.coop 2017).

Online platforms have been discussed as a way to facilitate large-scale data sharing between researchers, clinicians and patients (Tillmann et al. 2015). For example, the Open mHealth online platform aims to facilitate the sharing of open-source data analysis and visualisation tools, specifically for big health data for research purposes (Chen et al. 2012).

Another technological solution which is gaining momentum as a way to manage data storage securely includes distributed ledger technologies (DLTs) such as block chain (UK Government Chief Scientific Adviser 2016). A distributed ledger is a digital ledger in which, in contrast to centralised networks and ledger systems, information is stored on a network of machines, with changes to the ledger reflected simultaneously for all holders of the ledger and authenticated by a cryptographic signature, thus providing a transparent and verifiable record of transactions. In block-chain technology, the ledger comprises ‘blocks’ of transactions. One of the most well-known uses of DLTs is in the cryptocurrency Bitcoin. However, the possible uses of DLTs go well beyond the financial sector and include applications in education, the creative industries, agriculture and food, and health (Godsiff 2016; O’Dair 2016; Sharples & Domingue 2016).

Literature considering actions to support a data sharing health society also discusses dynamic consent models (Williams et al. 2015), through which patients are envisaged to realise additional levels of and opportunities for control over their data in a transparent manner, with the assumption that this would lead to greater public trust. These would involve patients being able to receive information on how the data is being used and consenting to various data sharing requests on a case-by-case basis (Williams et al. 2015), most probably through a user-friendly electronic model. Further research would need to consider how the capacity to manage these models of consent and the associated resource implications could be developed, as well as how such a model would relate to the GDPR and to efforts to simplify the system. In addition, any model relying on an electronic platform would need to be sensitive to unintended effects such as inability to engage individuals who are not computer literate or do not have access to a computer, potentially leading to inequities in inclusion. During our roundtable discussions, some participants highlighted practical challenges to implementing approaches which require direct patient consent for each use of a particular dataset. These individuals noted a need to explore novel approaches to governance and risk management which would fully prioritise patient autonomy (which cannot be compromised), but ensure that the governance of consent (responsibility and control) is delegated in a practically workable way (e.g. via intermediary and entrusted organisations and at national, regional or pan-EU levels). In these models patients would still have the right to prevent use of their individual data.

A particular literature discusses data sharing specifically in relation to clinical trials (Koening et al. 2015; Loder & Groves 2015; Mello et al. 2013). This is particularly relevant considering the October 2016 EMA decision to publish clinical data submitted by pharmaceutical companies to support their regulatory applications for
human medicines under the centralised procedure. Koening et al. (2015) discuss how sharing of clinical data should also consider the need to plan for workforce capacity in regulatory agencies to ensure sufficient oversight of data sharing practices, validate data accuracy, manage statistical and technical aspects of data sharing and access issues in a way that will ensure both commercial confidence and patient privacy. Calls to register and report results from all clinical trials – regardless of positive or negative findings – are also gaining increased attention (INT09; INT10; Koening et al. 2015).

4.4. Workforce capacity-building: supporting health professionals to make the most of health data

The various applications of new sets of health data can be seen as disruptive innovations that create new market categories as a result of discontinuous technology change or a disruptive business model (Corsham Institute and RAND Europe 2016). Healthcare professionals will need to diversify their skills and become more ‘technology-savvy’ in order to efficiently and effectively engage with a data-heavy healthcare delivery sector in the future (Wachter 2016). To support this, health systems will need to ensure appropriate training in both clinical informatics (Wachter 2016) and ‘softer skills’ needed to effectively access, interpret and communicate health data – both as part of medical education of future health and care professionals and as part of continual professional development programmes (Alyass et al. 2015; Hood & Auffray 2013; Mirnezami et al. 2012; Salas-Vega et al. 2015). Panhuis et al. (2014) highlight that such change will also require incentive mechanisms that reward new ways of working. As one interviewee stated: ‘A lot of the health professionals are very busy in their clinical work so they don’t have time to use as much as they could the data that is collected by other people.... Often this is not considered an important indicator of their work [the usage of data, at present]. It’s more important to see 200 people than to use the data collected from these 200 people. This is a challenge for the future. In evaluation there should be an indicator that the professionals have used the data that they collected’ (INT04). Concerns about legal liability related to decision-making that is informed by some types of health data (e.g. genomic information) and predictive analytics will also need to be considered.

Engaging health professionals in the planning, implementation and testing phases of different health data solutions can help in capacity-building to engage with data (Nguyen et al. 2014) as well as in nurturing supportive attitudes. Effective health workforce capacity-building in this space will depend on an appreciation of the value added that health data analytics can bring to delivery on job roles and responsibilities and simultaneously on empowering other stakeholders – such as patients and the public – to understand, appreciate and engage with health data ecosystems.

4.5. Learning from previous efforts to increase public awareness, acceptability and engagement with health data

Efforts to raise awareness on the potential of health data and on appropriate safeguards for its use need to be rooted in clear, well-targeted and comprehensive communication strategies (Hood & Auffray 2013; INT01–03; Keen et al. 2013; Kshetri 2014). For example, the challenges in achieving public acceptability and trust have been made explicit in the experience of trying to implement the care.data effort in England. This programme was established by NHS England and the Health and Social Care Information Centre (HSCIC – now NHS Digital) to integrate health and social care information from different healthcare settings (primary, secondary,
community care), so as to learn about and inform effective care delivery. However, care.data experienced significant challenges related to public acceptability. These related to concerns about data privacy, security safeguards and the consent process and were partially associated with the nature of the public communication and engagement campaign (NHS England 2017). Following a review led by Dame Fiona Caldicott (National Data Guardian 2016), the care.data programme was closed, but stakeholders remained committed to establishing a trusted data sharing system ultimately aimed at ensuring better patient outcomes and experiences through better-informed and evidence-based decision-making. The work initially started through care.data is being taken forward by the National Information Board, which brings together health and social care actors and organisations to set priorities and devise strategies for capturing value from IT and information. NHS Digital is closely involved with this work and with wider efforts to ensure that the public, health professionals, care commissioners and the research community can benefit from better access to health and care information (NHS Digital 2017).

Carter et al. (2015) identify learning from the care.data experience that would need to be acted on in similar national-scale initiatives in the future. Specific areas to address would include: (i) ensuring that communication media (leaflets in the case of care.data) provide sufficient explanations about the risks of data sharing, who will have access, the terms of data handling and opt-out forms (if applicable); (ii) consideration of the changes that would affect GPs and how the doctor–patient relationship would be affected in terms of new expectations and duties; and (iii) greater clarity on what data exists on patients and who has access to it and why, including the nature of the public good stemming from the programme (INT01, 03, 07; Kshetri 2014).

Participants in the expert roundtable for this study also highlighted that improving patients’ confidence requires not only a clear but also a consistent message on all issues of data usage from healthcare professionals, policymakers and the media. In the view of one interviewee: ‘Trust and agency are very important. People are extremely interested in contributing for their health and for the greater good of society. This potential is under-used. If the NHS had asked people to play a more active role in the care.data initiative they would have gained much greater acceptance’. (INT01) For example, the Scottish Health Informatics Programme (SHIP), a research platform for the collecting, dissemination and analysis of anonymised electronic patient records, has involved different stakeholders – including the public and patients – in efforts to develop mechanisms of principled proportionate governance. This has been done through public engagement exercises to understand how best to manage risks associated with health data use in research. The intensive public engagement is seen as a key success factor for SHIP (Carter et al. 2015).

Auffray et al. (2016), Hood and Auffray (2013) and Williams et al. (2015) also suggest that public and patient engagement through social media could be a potential channel for communicating accessible information. One interviewee highlighted the importance of pharmaceutical industry participation and initiative in engaging the wider public approaches: ‘...pharmaceutical companies need to...make the data experience for patients extremely engaging and not only to deal with the data needs of the company but to put [these] data needs...in the context of a much broader patient journey, and to have a vision that what the patient starts doing when they are ill will actually continue to be engaging and useful for them as citizens after they have stopped being ill...’ (INT05).

Public communication efforts should also clarify safeguards that would protect the public in regard to asymmetric information that may lead to adverse selection or moral hazard (e.g. in health insurance). For example a campaign for equitable sharing of risks and benefits should address fears of discrimination against
people living in areas with a high incidence of certain diseases who could suffer from increasing insurance premiums (King 2014).

Respecting legal provisions and patients’ rights in the context of using health data for research purposes is a topic of increasing attention and focus in recent times (EC 2016; GDPR 2016; Salas-Vega et al. 2015), although current directives lack optimal clarity on eligibility and boundaries to what constitutes research. Cohen et al. (2014) discuss various ways of meeting these needs, including: secure and trusted de-identification; privacy safeguards such as providing free credit violation monitoring; tort remedies (damages to compensate for harm inflicted) for patients who suffer harm due to privacy breaches; seeking additional levels of consent; introducing certification processes for people or institutions using certain types of health data; and ‘naming and shaming’ strategies that publicise serious data breaches and the actions that are being taken.

Finally, roundtable participants highlighted the need for a more granular approach to understanding patient and public acceptability of data sharing and use. The desired approach would consider trade-offs associated with different types of data use, as well as how public attitudes related to distinct types of data.

4.6. In reflection

Health data and its numerous applications are forming a highly dynamic field that has been developing in recent years. As our analysis has shown, the potential benefits that health data-driven sectors can have for research and innovation landscapes, ensuring drug safety and public health, healthcare delivery and health systems more widely, are great. Demonstrating value for money from health data initiatives will inevitably play a key role in informing future investments. Participants in the expert roundtable for this research also highlighted a need to consider upfront costs and decommissioning needs associated with health data ecosystems alongside short-, medium- and longer-term benefits. However, achieving and demonstrating value for money from health data initiatives is challenging, in particular due to a lack of counterfactual and the time needed for financial benefits to materialise (Wachter 2016).

There is a need for more learning-focused evaluation to inform interventions at system levels. Evaluations of existing efforts and individual initiatives do take place (although not systematically) but tend to focus more on capturing outputs than on identifying lessons for wider-scale spread, scale-up and sustainability of interventions across contexts. Existing learning has also not been brought together and shared as well as it could be, to enable transferable insights and a critical mass of evidence on the cost-effectiveness and value for money, scalability potential and sustainability of comparable types of health data efforts and institutional initiatives. Such evaluation could also offer meaningful learning to inform future initiatives in terms of issues such as risk management and implementation resource and capacity needs. Both retrospective and real-time approaches, qualitative and quantitative, would form part of a mix that could support progress. Discrete choice modelling could help inform about the trade-offs individuals, stakeholder groups and nations are willing to make, under alternative scenarios. In addition, prospective economic analyses and modelling could also help inform the contributions of a data-driven health sector to economic development and international competitiveness.

As this study has highlighted, realising the benefits of health data at scale will require a focus not only on (i) creating the technological and structural conditions needed for data capture, curation and transformation into high-quality and actionable evidence, but also on (ii) timely, collaborative and coordinated efforts to transform working cultures, build health and care workforce and citizen capacity to engage with data, and on (iii) policy,
industry and research community responsiveness to public concerns, needs and expectations – in order to create and connect receptive places for health data. As a European and global community, consideration should also be given to equitable progress across different geographical contexts which may be at different stages of development of their health data infrastructures.
References


DZHK. 2013. ‘Gutenberg Health Study is Supported by Boehringer Ingelheim for Another Four Years.’ Dzhk.de, 29 August. As of 6 March 2017: https://dzhk.de/en/news/latest-news/article/gutenberg-health-study-is-supported-by-boehringer-ingelheim-for-another-four-years/


Genomics England. 2017. ‘Research in the 100,000 Genomes Project.’ Genomicsengland.co.uk. As of 17 March 2017: https://www.genomicsengland.co.uk/about-gecip/


Ludman, P.F. 2012. ‘The UK Transcatheter Aortic Valve Implantation Registry; One of the Suite of Registries Hosted by the National Institute for Cardiovascular Outcomes Research (NICOR).’ *Heart* (British Cardiac Society) 98(24).


Maudsley Biomedical Research Centre Dementia Biomedical Research Unit. 2017. ‘Clinical Record Interactive Search.’ Maudsleybrc.nihr.ac.uk. As of 17 March 2017: http://www.maudsleybrc.nihr.ac.uk/about-us/core-facilities/clinical-record-interactive-search-cris/


———. 2016. ‘Dataiku and Bioserenity Team Up to Develop Wearable Device for Improving Epilepsy Diagnosis.’ News-medical.net, 2 September. As of 17 February 2017: http://www.news-


University of Bristol. 2015. ‘IRC-sphere.’ Irc-sphere.ac.uk. As of 17 March 2017: http://www.irc-sphere.ac.uk/
University of Edinburgh. 2017. ‘Research Case Study: Good Governance for the Scottish Health Informatics Programme (SHIP).’ Law.ed.ac.uk. As of 17 March 2017: http://www.law.ed.ac.uk/research/making_a_difference/ship


Appendix A Methodological Approach

In November 2016 we searched Google Scholar using a combination of the following terms: health, patient data, clinician data, provider data, health systems data, real-world data, big data, open data, value, benefits, impacts, challenges and potential. We used a limited geographical focus for the European countries. Articles in English published from 2010 onwards were considered. The search yielded 6,350 results. The first 350 articles were screened for relevance based on their title and abstracts. One researcher performed the first screening and introduced the titles in an Excel tracker with reasons why some were excluded. A second researcher reviewed the tracker. Discussions on a potential list of includable studies took place between the two until consensus was reached. In total 107 papers were identified as potentially relevant. For these articles the tracker was updated with the publication data and number of citations of each article as well as key words on their content. The researchers further narrowed the number of identified articles to 40 based on a combination of criteria including number of citations, date of publication (the most recent were sought) and an interest in balanced coverage of relevant topic areas (avoiding selecting a disproportionate number of articles on the same topic at the expense of others on a different but relevant topic).

The first set of scholarly literature papers (i.e. journal articles) was complemented with an additional 17 resulting from grey literature sources – either normal Google search or websites and initiatives that were recommended during the interview process. The majority of selected papers represented reviews of existing literature, general overviews of developments with examples of initiatives meant to capture the value of health data, and opinion pieces. There were very few papers that used primary data collection methods (King et al. 2012 – interviews; EC 2016 – Delphi method). The selected literature is dominated by authors from high-income countries such as the UK, the United States, the Netherlands and Canada.

Following a snowballing technique from the selected articles as well as recommendations from the interviewees an additional 92 sources of information (scholarly literature, reports, web sites) were consulted. All of these were read in full text but were not subjected to the systematic data extraction used on the previous 57 sources.

The 12 interviews were conducted in January and February 2017 and involved discussions with the following types of stakeholders: 3 representatives of Academia, Research Institutes, Research Councils and Think Tanks, 3 of Consumer and Patients Associations, 3 of Policy/Government/EC, 3 of Private Sector, Trade Associations or Regulatory Bodies or Other. Interviews were semi-structured, using an interview protocol, and lasted between 40 and 60 minutes. Informed consent forms were obtained from participants prior to the interview. Due to the small sample size of interviewees we cannot expect to have reached saturation for the views of each stakeholder group. The interviews served to test and enrich the findings from the core aspect of this study – namely the literature review – and inform areas for potential future consideration.

The findings from the report were discussed during a roundtable event on 17 February 2017. The event benefited from input from 22 participants (not including RAND Europe researchers) and consisted of a series of
workshop sessions discussing various examples of European initiatives that make use of health data, as well as advancements and ways to tackle issues such as data protection, public awareness and collaboration and coordination in the health data ecosystem/data access and sharing. Reflections from the workshops were subsequently incorporated in the present report.
## Appendix B List of Consulted Experts

<table>
<thead>
<tr>
<th>Name</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brendan Barnes</td>
<td>Director, Data Protection, IP and Global Health, European Federation of Pharmaceutical Industries and Associations</td>
</tr>
<tr>
<td>Michael Berntgen</td>
<td>Head of Scientific and Regulatory Management Department, European Medicines Agency</td>
</tr>
<tr>
<td>Hassan Chaudhury</td>
<td>Founding Director and Chief Commercial Officer, Health IQ</td>
</tr>
<tr>
<td>Ludivine Blanc</td>
<td>Manager, European Federation of Pharmaceutical Industries and Associations</td>
</tr>
<tr>
<td>Lucy Davis</td>
<td>Patient, Melanoma Patient Network Europe</td>
</tr>
<tr>
<td>Kevin John Dean</td>
<td>Managing Director, Smart Health Science</td>
</tr>
<tr>
<td>Jon Fistein</td>
<td>Director, Fistein Enterprise</td>
</tr>
<tr>
<td>Luciano Floridi</td>
<td>Professor of Philosophy and Ethics of Information, Oxford Internet Institute, University of Oxford</td>
</tr>
<tr>
<td>Francesco Florindi</td>
<td>Head of EU Affairs, European Cancer Patient Coalition</td>
</tr>
<tr>
<td>Ernst Hafen</td>
<td>Deputy Head of Institute for Molecular Systems Biology, ETH Zurich</td>
</tr>
<tr>
<td>Shahid Hanif</td>
<td>Head of Health Data &amp; Outcomes, Association of the British Pharmaceutical Industry</td>
</tr>
<tr>
<td>Stanimir Hasardzhiev</td>
<td>Founder and Chairperson, National Patients’ Organisation (Bulgaria)</td>
</tr>
<tr>
<td>Niklas Hedberg</td>
<td>Chief Pharmacist, Dental and Pharmaceutical Benefits Agency (Sweden)</td>
</tr>
<tr>
<td>Dipak Kalra</td>
<td>Clinical Professor of Health Informatics at University College London and Director of the Centre for Health Informatics and Multiprofessional Education</td>
</tr>
<tr>
<td>Dieter Kugelmann</td>
<td>Data Protection Commissioner, German Federal State, Rhineland Palatinate</td>
</tr>
<tr>
<td>Marialuisa Lavitrano</td>
<td>Prorector for International Affairs at the Milano-Bicocca University and co-director of Biobanking and BioMolecular resources Research Infrastructure</td>
</tr>
<tr>
<td>Robert Madelin</td>
<td>Hors Classe Adviser ”Senior Adviser for Innovation” in the European Political Strategy Centre (EPSC), former Director General of DG Connect and DG Sanco</td>
</tr>
<tr>
<td>Francesco Modafferi</td>
<td>Head of Inspection Department, Italian Data Protection Authority</td>
</tr>
<tr>
<td>Nicola Perrin</td>
<td>Head of Policy, Wellcome Trust</td>
</tr>
<tr>
<td>Raminderpal Singh</td>
<td>Vice President, Business Development, Eagle Genomics</td>
</tr>
<tr>
<td>Paul Timmers</td>
<td>Former Director of Digital Society, Trust and Security in the EC Communications Networks, Content and Technologies Directorate General</td>
</tr>
<tr>
<td>Antoni Sicras</td>
<td>Scientific Director of the Rediss Foundation</td>
</tr>
<tr>
<td>Alexander Whalen</td>
<td>Senior Policy Manager, Digital Economy Policy Group Digitaleurope</td>
</tr>
<tr>
<td>Petra Wilson</td>
<td>Senior Advisor at FTI Consulting EU and Managing Director of Health Connect Partners</td>
</tr>
</tbody>
</table>

## Appendix C Value of Health Data – Selected Examples

<table>
<thead>
<tr>
<th>Case study</th>
<th>Status</th>
<th>Description</th>
<th>Type of data covered</th>
<th>Envisaged benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achmea Health Database (UMC Julius Center, 2017)</td>
<td>Ongoing</td>
<td>A registry by the health insurance company AGIS that contains all healthcare procedures carried out in the last ten years using the records of 1.2 million health-insured individuals in the Netherlands.</td>
<td>Claims data; real-world data</td>
<td>Could open opportunities for data mining to support research and could improve knowledge about healthcare consumption to evaluate quality of care and inform service planning, provided regulation is in place to ensure appropriate use.</td>
</tr>
<tr>
<td>ADR-PRISM Project (ADR-PRISM, 2017)</td>
<td>Ongoing</td>
<td>A project that aims to leverage data from patient discussion forums for pharmacovigilance professionals.</td>
<td>Citizen-generated social media data</td>
<td>Could advance scientific and medical knowledge, which could lead to better health outcomes.</td>
</tr>
<tr>
<td>BioSerenity and Dataiku’s NEURONAUTE (News-Medical, 2017)</td>
<td>Ongoing</td>
<td>A wearable device that aids in diagnosis and real-time monitoring of epilepsy.</td>
<td>Patient-generated mobile apps, telemedicine and sensor data</td>
<td>Could allow doctors and care providers more access to information, thereby allowing quicker reaction rates and potentially improved care outcomes.</td>
</tr>
<tr>
<td>BMJ Open Data Campaign (BMJ, 2017)</td>
<td>Ongoing</td>
<td>A BMJ policy that requires publication in the BMJ to be predicated by access to anonymised patient-level data underpinning studies on trials of drugs or devices.</td>
<td>Applies to any form of sensitive health data / patient-level data</td>
<td>Could promote data transparency and data sharing.</td>
</tr>
<tr>
<td>The Clinical Practice Research Datalink (CPRD, 2017)</td>
<td>Ongoing</td>
<td>An English NHS, not-for-profit research service that is jointly funded by the NHS NIHR and the Medicines and Healthcare products Regulatory Agency of the UK Department of Health, which provides anonymised primary care records for public health research.</td>
<td>Healthcare professionals-provided EHR data</td>
<td>Makes anonymised data available for researchers, who could then use the data for research that could lead to improvements in drug safety, best practice and clinical guidelines.</td>
</tr>
</tbody>
</table>
| Cloud computing at London’s Chelsea and Westminster Hospital (Sultan, 2014) | Ongoing | A cloud-based electronic medical records system that holds patient data in a centralised system and can make patient data available with the patient’s consent to hospital staff, GPs and trusted family members (Toor, 2011).                                                                                                                                                                                                                                                                                     | Healthcare professionals-provided EHR data               | Could enable easy access to medical information and could empower patients to have more control over their treatment. Could give health professionals easy access to wide range of information that can underline trends in how patients...
| **Hungary’s national health app**  
*Collis, 2016* | Ongoing | A mobile health app that allows patients access to a cloud system with their EHR from hospitals and community practices. Patients voluntarily upload information about their diet, weight, blood pressure, glucose levels, exercise and conditions such as pregnancy, as some examples. The app then provides tailored lifestyle and treatment advice to individuals, and individuals can receive advice from their GP. Patients can also control which health workers they allow to see their data. | Healthcare professionals-provided EHR data and patient-provided mobile apps data | Aims to improve public health, with special focus on people with chronic diseases.  
Aims to generate cost-savings through preventative care, potentially reducing costs associated with more reactive measures.  
Aims to facilitate improved health outcomes through patients’ self-management and the app’s facilitation of evidence-based decision-making for the GP. |
| **Ensuring Consent and Revocation Project**  
*EnCoRe*  
*EnCoRe, 2017* | Ongoing | A research project by UK industry and academia working to develop consent management technologies, IT infrastructures that allow for informed consent and the regulatory governance necessary to enhance and strengthen health data systems. | Applies to any type of sensitive health data | Could empower patients to have more control over their personal information and how it is used. |
| **EU Adverse Drug Reaction (EU-ADR) Web Platform**  
*EU-ADR, 2017* | Completed  
(2008–2012) | An EC project that aimed to exploit existing electronic healthcare records for drug safety alerts based on patterns of association between treatments and symptoms experienced. | Healthcare professionals-provided EHR data | Aimed to improve knowledge on adverse drug reactions for pharma, physicians and users to improve drug safety and contribute to public health. |
| **European Health Records for Clinical Research (EHR4CR) project**  
*EHR4CR, 2017* | Completed  
(2011–2016) | An EU-wide technological platform that securely connects hospital EHR systems and clinical data warehouses across Europe, and uses de-identified patient data to assist in clinical trials’ feasibility assessments and patient recruitment. | Healthcare professionals-provided EHR data | Aimed to equip clinical trial sponsors with predictive analytics on the number of eligible patients for particular trials to help identify appropriate trial sites as well as create a channel connecting trial sponsors with providers able to assist the recruitment process. |
<p>| <strong>European Medical Information Framework (EMIF) Data Catalogue</strong> | Ongoing | A platform that aims to integrate data as well as carry out data source profiling, which could allow researchers to browse metadata and could provide access to existing health data repositories such as cohort studies, biobanks and EHR to researchers, thereby potentially improving identification, access and | Healthcare professionals-provided EHR data | Could extend access to existing health data repositories |</p>
<table>
<thead>
<tr>
<th>Project/Database</th>
<th>Status</th>
<th>Description</th>
<th>Type of Data</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>EuResist Integrated DataBase (EIDB) (Miani et al. 2014)</td>
<td>Ongoing</td>
<td>A project that builds on the collaboration between pharmaceutical companies, governmental institutions, private companies and other partners. The project’s new technologies and mathematical models aim to provide a more efficient way to choose the best drugs and drug combinations for any given HIV genetic variant by predicting responses to treatment. It aims to create an online system that helps doctors to choose a HIV treatment with the highest probability of halting virus replication and impairing the evolution of drug resistance.</td>
<td>Research-generated omics data (genotype data)</td>
<td>To provide more personalised medicine and more effective treatment outcomes.</td>
</tr>
<tr>
<td>Genome Medicine Centres and Genomics England Clinical Interpretation Partnerships (GeCIPs)</td>
<td>Ongoing</td>
<td>A research network that aims to help enrol participants into Genomics England, accept NHS referrals and ensure appropriate sample and phenotypic information collection. As part of this initiative, genomic data will be related to phenotypic research data and clinical records to form a research dataset – for example to inform drug development, as well as to provide feedback to the individual patients via their clinicians.</td>
<td>Research-generated omics data (genotype data)</td>
<td>Could generate research to enhance drug development, that will potentially benefit patients via improved diagnosis and treatment effectiveness (Dubow and Marjanovic, 2015).</td>
</tr>
<tr>
<td>GSK Clinical Study Register (GSK, 2017)</td>
<td>Ongoing</td>
<td>A data repository with protocol summaries, scientific results summaries, clinical study reports and more from GSK-sponsored clinical trials.</td>
<td>Research-generated clinical-trial data</td>
<td>Could improve transparency of clinical trial data.</td>
</tr>
<tr>
<td>General Practice Research Database (HERC, 2017)</td>
<td>Completed</td>
<td>A database of anonymised and longitudinal primary-care EHR jointly funded by the NHS NIHR and the Medicines Healthcare products Regulatory Agency.</td>
<td>Healthcare professionals-provided EHR data</td>
<td>Aimed to allow academics, government departments, pharma companies and the NHS (MHRA, 2017) access to large volumes of anonymised data from patients’ medical records in order to investigate effects of medicines and outcomes of treatments, as well as to evaluate which medicines work best (MHRA, 2017).</td>
</tr>
<tr>
<td>Gutenberg Health Study</td>
<td>Ongoing</td>
<td>Started in 2007 the study aims to establish a biobank</td>
<td>Research-generated data</td>
<td>These advancements could lead to improved diagnosis and</td>
</tr>
<tr>
<td>Project</td>
<td>Status</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(DZHK, 2013)</td>
<td></td>
<td>Based on the collected data. Expected contributions are in the area of predicting individuals’ risk of the development of cardiovascular disease, cancer, eye diseases, metabolic diseases, disorders of the immune system and the psyche (DZHK, 2013).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IBM’s Shared Care platform (News-Medical, 2013)</td>
<td>Ongoing</td>
<td>A project based in southern Denmark that involves two hospitals, three municipalities and several medical practices, as well as IBM. The main goal is to facilitate real-time communication among patients, physicians, pharmacists and mental health professionals, so all parties have insight over patients’ status. It uses large volumes of data from the clinical setting, analyses the data to generate insights for healthcare providers, and integrates health records among health, social services and other providers. The programme initially focused on heart disease and will later include patients with type II diabetes and pulmonary conditions.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IBM ‘Watson’ Health (with application to health analytics) (IBM, 2017)</td>
<td>Ongoing</td>
<td>A computer system that uses ‘cognitive computing’ developed for analysing big data that puts large volumes of information such as medical literature and EHR into context, exploiting text information from publicly available sources.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Innovative Medicine Initiative’s Remote Assessment of Disease</td>
<td>Ongoing</td>
<td>A project that involves the collaboration of diverse stakeholders and sectors such (academics, healthcare professionals, industry, patients and citizens, clinical Mobile, sensor and wearable technology data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Therapy. A partnership between the University Medical Center of the Johannes Gutenberg-University Mainz and Boehringer Ingelheim, the project is unique in Germany and Europe due to its big sample size as well as the holistic view of diseases involving different expositions. A workshop participant also explained that the study has managed to resolve consent issues encountered throughout the project and has increased research capacity at local level.
<table>
<thead>
<tr>
<th><strong>and Relapse in Central Nervous System Disorders (RADAR-CNS)</strong></th>
<th>service providers, payers and regulators across digital and life-science sectors (Wicks et al. 2016). The programme is co-led by Janssen and King’s College London and aims to develop new means of monitoring major depressive disorder, epilepsy and multiple sclerosis with the help of wearable devices and smartphone technology (Innovative Medicines Initiative, 2017).</th>
<th>data for research that could lead to medical breakthroughs.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The Maudsley Biomedical Research Centre’s Clinical Record Interactive Search system (Maudsley Biomedical Research Centre Dementia Biomedical Research Unit, 2017)</strong></td>
<td>A system that has been developed for use within the NIHR Maudsley Biomedical Research Centre and Dementia Unit, that uses de-identified electronic medical records data.</td>
<td>Healthcare professionals-provided EHR data</td>
</tr>
<tr>
<td><strong>Midata.coop (MIDATA, 2017)</strong></td>
<td>An initiative that enables citizens to store their information in one secure place in a not-for-profit cooperative model. The citizens can decide whether they wish to share this data with healthcare providers or to participate in research and grant access to subsets of their data.</td>
<td>Citizen-provided data</td>
</tr>
<tr>
<td><strong>NHS Atlas of Healthcare Variation (Exeter et al. 2014)</strong></td>
<td>A project using interactive mapping software that can be used to map select indicators of interest, in order to detect geographical determinants of health such as deprivation, lifestyle and environmental factors.</td>
<td>Geospatial health data (deprivation, lifestyle and environment data)</td>
</tr>
<tr>
<td><strong>Open mHealth (Open mHealth, 2017)</strong></td>
<td>An online platform that aims to facilitate the sharing of open-source data analysis and visualisation tools, specifically for big health data. Independently</td>
<td>Data analysis tools</td>
</tr>
<tr>
<td><strong>Midata.coop (MIDATA, 2017)</strong></td>
<td>An initiative that enables citizens to store their information in one secure place in a not-for-profit cooperative model. The citizens can decide whether they wish to share this data with healthcare providers or to participate in research and grant access to subsets of their data.</td>
<td>Citizen-provided data</td>
</tr>
<tr>
<td><strong>NHS Atlas of Healthcare Variation (Exeter et al. 2014)</strong></td>
<td>A project using interactive mapping software that can be used to map select indicators of interest, in order to detect geographical determinants of health such as deprivation, lifestyle and environmental factors.</td>
<td>Geospatial health data (deprivation, lifestyle and environment data)</td>
</tr>
<tr>
<td><strong>Open mHealth (Open mHealth, 2017)</strong></td>
<td>An online platform that aims to facilitate the sharing of open-source data analysis and visualisation tools, specifically for big health data. Independently</td>
<td>Data analysis tools</td>
</tr>
<tr>
<td><strong>Midata.coop (MIDATA, 2017)</strong></td>
<td>An initiative that enables citizens to store their information in one secure place in a not-for-profit cooperative model. The citizens can decide whether they wish to share this data with healthcare providers or to participate in research and grant access to subsets of their data.</td>
<td>Citizen-provided data</td>
</tr>
<tr>
<td><strong>NHS Atlas of Healthcare Variation (Exeter et al. 2014)</strong></td>
<td>A project using interactive mapping software that can be used to map select indicators of interest, in order to detect geographical determinants of health such as deprivation, lifestyle and environmental factors.</td>
<td>Geospatial health data (deprivation, lifestyle and environment data)</td>
</tr>
<tr>
<td><strong>Open mHealth (Open mHealth, 2017)</strong></td>
<td>An online platform that aims to facilitate the sharing of open-source data analysis and visualisation tools, specifically for big health data. Independently</td>
<td>Data analysis tools</td>
</tr>
<tr>
<td><strong>Midata.coop (MIDATA, 2017)</strong></td>
<td>An initiative that enables citizens to store their information in one secure place in a not-for-profit cooperative model. The citizens can decide whether they wish to share this data with healthcare providers or to participate in research and grant access to subsets of their data.</td>
<td>Citizen-provided data</td>
</tr>
<tr>
<td><strong>NHS Atlas of Healthcare Variation (Exeter et al. 2014)</strong></td>
<td>A project using interactive mapping software that can be used to map select indicators of interest, in order to detect geographical determinants of health such as deprivation, lifestyle and environmental factors.</td>
<td>Geospatial health data (deprivation, lifestyle and environment data)</td>
</tr>
<tr>
<td><strong>Open mHealth (Open mHealth, 2017)</strong></td>
<td>An online platform that aims to facilitate the sharing of open-source data analysis and visualisation tools, specifically for big health data. Independently</td>
<td>Data analysis tools</td>
</tr>
</tbody>
</table>
| **Optimum Patient Care Research Database (OPCRD)**  
(OPCRD, 2017) | Ongoing | A database that contains data on patient-reported experience and anonymised longitudinal medical records data from over 550 UK general practices, as well as the results of patient-completed questionnaires. | Healthcare professionals-provided EHR data and patient-provided data | Could enable researchers to mine data, which could lead to new treatments and breakthroughs, thereby leading to improved health outcomes. For example, using patient-completed questionnaires from the OPCRD, a research team that included researchers from GSK and Novartis investigated the relative efficacy of drugs used to treat COPD and concluded that COPD is not treated according to the recommendations of the Global Initiative for Chronic Obstructive Lung Disease and the NICE (Price et al. 2014). |
| **Patients Know Best**  
(Patients Know Best, 2017) | Ongoing | A system that stores information in a secure way and allows patients to choose whom they want to share their data with. The portal uses the first Consent Engine in a Patient Portal with the records being categorised into different disciplines that can later be shared depending on patient consent (Patients Know Best, 2017). | EHR; real-world data; | Greater patient engagement, with Patients Know Best reporting over 70 per cent uptake of the platform compared to less than 20 per cent for other patient portals. Patients Know Best also reported: a 30 per cent reduction in unplanned GP visits post-surgery; better diagnosis when it comes to epilepsy services; and reduction of unplanned emergency admissions for long-term conditions. Savings to the system are mentioned in the range of £3 to £5 for every £1 spent on Patients Know Best. These are a result of going paperless at the point of care or savings from costs associated with non-attendance rates, reducing the time spent on telephone calls for clinicians, setting up unnecessary appointments or undertaking administrative tasks such as printing and photocopying (Patients Know Best, 2017). |
| **PatientsLikeMe® Epilepsy Community Platform**  
(Completed) | Completed | A subsection of the PatientsLikeMe portal established in January 2010 in the United States in partnership with UCB Pharma including its European branches, where patients with epilepsy can record, monitor and share their data for the purposes of this specific trial. | Patient-generated social network data (demographic, disease and treatments) | Aimed to enable a greater understanding of epilepsy patient perceptions about the effects of their treatments and their quality of life (de la Loge et al. 2016). Based on retrospective analysis, the study revealed a high rate of key symptoms such as memory problems, fatigue and problems concentrating, as well as poor quality of life.
<table>
<thead>
<tr>
<th>Project Name</th>
<th>Status</th>
<th>Description</th>
<th>Data Type</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Safety Of Non-Steroidal Anti-Inflammatory Drugs (SOS) project</td>
<td>Completed</td>
<td>An EU FP7 project that investigated drug safety for NSAIDs – in particular the association between these and gastro-intestinal and cardiovascular effects of traditional NSAIDs and COX-II selective NSAIDs (Woerdeman et al. 2017). Meta-analysis had shown that clinical trials were not large enough to show convincing safety estimations for gastro-intestinal and cardiovascular effects, particularly in children, while meta-analysis from observational studies did not provide estimates for heart failure. Linking data from 8.5 million new NSAID users, the SOS database study was able to inform a decision model for choosing the least toxic NSAID in respect to gastro-intestinal and cardiovascular effects (Woerdeman et al. 2017).</td>
<td>Real-world data</td>
<td>Could aid clinical decision making and potentially lead to improved care outcomes.</td>
</tr>
<tr>
<td>Salford Lung Study (University of Manchester, 2016)</td>
<td>Completed</td>
<td>An innovative clinical trial design that used software to integrate EHR of patients’ interactions with doctors, pharmacists and hospitals to investigate the effectiveness and safety of a new treatment for COPD.</td>
<td>Healthcare professionals-provided EHR data; real-world data</td>
<td>Aims to allow researchers to access a broader and more inclusive population of patients that would include patients who would usually be excluded by conventional RCTs’ stringent inclusion criteria (such as patients who are simultaneously treated for other chronic conditions), which could then potentially allow researchers to test treatments in a more realistic setting.</td>
</tr>
<tr>
<td>Scottish Health Informatics Programme (SHIP) Information Governance Toolkit</td>
<td>Ongoing</td>
<td>The Scottish Health Informatics Program (SHIP) is a research platform for collaboration between Scottish universities and NHS Scotland that collates, disseminates and analyses EHR. The SHIP toolkit, through public engagement exercises, is an initiative that seeks to ensure that ethical, legal, social and practical issues are identified (Sethie and Laurie, 2017).</td>
<td>Healthcare professionals-provided EHR data</td>
<td>Aims to develop mechanisms of proportionate governance that ensure that ethical and legal considerations regarding the secondary use of patients’ data are adhered to (SHIP, 2017). Ultimately, the initiative also aims to ensure that data is used in the public interest (University of Edinburgh, 2017). The initiative has been commended for its effective patient and public involvement campaigns.</td>
</tr>
<tr>
<td>Project Name</td>
<td>Status</td>
<td>Description</td>
<td>Data Use Case</td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------------</td>
<td>--------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Single-shared assessment system (SSA)</td>
<td>Ongoing</td>
<td>The project implies completing a Single Shared Assessment (SSA) form that is designed to assess the patient’s community care and healthcare needs by gathering data such as medical history, home environment and current medication (Inverclyde Council, 2017). This data sharing system in Scotland may share the individual’s data with other relevant agencies with the permission of the patient (Aberdeen City Council, 2017). It is a system of data sharing that unifies previously separate systems in finance, management and delivery of services and shares electronic records among people who use the services, the agencies and the professionals who deliver the service (King et al. 2012).</td>
<td>Could streamline and speed up the process of assessing an individual’s need for services by helping agencies responsible for the health, social and housing needs of the individual coordinate with each other.</td>
<td></td>
</tr>
<tr>
<td>Social Network Enabled Flu Trends (SNEFT) Framework (Chen, 2017)</td>
<td>Completed (2010–2011) – US initiative but identified during review of EU studies in scope of this literature review</td>
<td>A data collection framework that continuously monitored online data from the social media platform Twitter relating to influenza-like illness epidemics.</td>
<td>Citizen-reported social media data. Aimed to provide large volumes of information that could help in early warning and prevention against influenza-like illnesses such as H1N1.</td>
<td></td>
</tr>
</tbody>
</table>
| Sphere project in the UK                          | Ongoing      | Funded by the Engineering and Physical Sciences | Real-world data | Results from the project could help diagnose and manage
| Strategic Intellectual Property Insight Platform (SIIP) Database | Ongoing | A cloud-based database designed for collating global patent data and scientific literature, which contains over 30 million documents and over 200 million annotated chemical compounds, enabling insight into complex chemical and biological patterns. | Patents and intellectual property data and researcher-provided data on chemical and biological compounds | Aims to improve R&D productivity, help users identify potential collaborators or acquisition targets, generate competitive intelligence and increase likelihood of successful bids for patents. |
| Structural Genomics Consortium (SGC) | Ongoing | A partnership between public, private and third-sector actors, which aims to use biomedical data in an innovative way and aims to cost-effectively produce 3D protein structures on a large scale. Through an open-access scheme the Consortium is placing all the novel protein structures in a Protein Data Bank and also sends samples to researchers on request for scientific purposes. In its first ten years of existence, the SGC has discovered and published approximately 40 chemical compounds and produced 452 peer-review journals and eight books (Morgan-Omics data (protein data) | The SGC model is also thought to lead to cost-savings with Biohub estimating in 2010 that the SGC’s efforts into the effectiveness of available antibodies will lead to over $750m in savings (Morgan-Jones et al. 2014). |
| The TRANSFoRm Project (I-HD, 2017) | Completed (2010–2015) | An EC FP7-sponsored project that sought to develop the core technology and ICT infrastructure that will build a ‘learning healthcare system’. The aim was to build a core digital information system to support the integration of research and clinical data. | Healthcare professionals-provided EHR data | Aimed to enable interoperability between clinical and research systems for specific aspects of research such as feasibility assessment and subject identification for participation in research. Aimed to allow researchers to search aggregated data across different sources and countries. |
| United Kingdom Transcatheter Aortic Valve Implantation (TAVI) Registry (Ludman, 2012) | Ongoing | The registry collects data from all hospitals in the UK on all TAVI procedures conducted in the UK since the technique was first used in 2007. | Healthcare professionals-provided data; real-world data | Makes it possible for researchers to access data in order to potentially improve the quality of the TAVI procedure and improve health outcomes. |
| Vaccine Adverse Event Surveillance and Communication (VAESCO) (ECDC, 2012) | Completed (ended in 2012) | A research network funded by the European Centre for Disease Prevention and Control, which aims to establish a collaborative network of regulatory agencies, public health institutes and academia able to collect and aggregate information on adverse vaccine events in Europe through a shared research infrastructure and collaborative vaccine safety studies. For example, the VAESCO consortium conducted a study on narcolepsy using data from electronic medical records in multiple European countries. | Healthcare professionals-provided EHR data | Aimed to facilitate data sharing of research findings and could contribute to expansion of scientific and medical knowledge. |
In addition the following list of examples were mentioned during the workshop:

- IC PerMed: EU H2020 project – International Consortium for Personalised Medicine
- ERN – European Reference Networks for rare diseases
- Decipher project based at EMBL-EBI
- SysVasc
- ESFRI roadmap includes several relevant infrastructures, e.g.
- IMI projects (various)
- Global Alliance for Genomics and Health (GAGH)
- Italian prevention database ‘Genomics in the Health System’
- Longitudinal population studies in the UK
- Disease registries in UK (cancer, cystic fibrosis, multiple sclerosis). The cystic fibrosis registry achieved impact through collaboration with industry, which led to accelerated innovation and the development of a treatment; the multiple sclerosis registry helped increase focus on symptoms and patient needs.
- National Cancer Institute: created match programme – standard molecular profile for tumours
- CDR initiative (Yoda, MSD)
- TCGA-ICGC – research repositories for genomics data on research patients – for drug development. Involves academia and industry. Cited by group member as having impact.
- National clinical audits – having impact on care.
- ‘Feet-first’ campaigns for diabetes – these also have clinical relevance (related to foot ulcers).
- Denmark had e-health records fully online (very detailed) but only visible to patients and prescribers; in contrast, Sweden had a large, widely accessible database but with very little detail – like Snapchat vs. a phonebook. Both have pluses/minuses.
- Gutenberg Health Study
- Sphere project in the UK – funded by the EPRSC, the project is investigating and developing the use of various sensors for the home
- Microsoft Alliance for Cloud Computing for Healthcare
- Patients Know Best
- Cancer registry, Germany
- Italy – BBMRI Expert Centre
- Denmark: individual data on web (individual EHR)
- Sweden: big database on drug packages used in ambulatory care, linkages possibly anonymised, big data analysis of drug consumption but no knowledge on outcomes.