Our data, our society, our health: a vision for inclusive and transparent health data science in the UK and beyond

Article (Accepted Version)

Our data, our society, our health: A vision for inclusive and transparent health data science in the UK and Beyond

A position statement

Authors:

Elizabeth Ford[1], Andy Boyd[2], Juliana K. F. Bowles[3], Alys Havard[4], Robert W. Aldridge[5], Vasa Curcin[6], Michelle Greiver[7], Katie Harron[8], Vittal Katikireddi[9], Sarah E. Rodgers[10,11], and Matthew Sperrin[12]

1. Department of Primary Care and Public Health, Brighton and Sussex Medical School, Brighton, UK.

2. ALSPAC, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK.

3. School of Computer Science, University of St Andrews, St Andrews, Scotland, UK.

4. Centre for Big Data Research in Health, UNSW Sydney, NSW 2052, Australia.

5. Institute of Health Informatics, UCL, London UK.

6. School of Population and Environmental Health Sciences, Faculty of Life Sciences and Medicine, King's College London, UK.

7. Department of Family and Community Medicine, University of Toronto, North York General Hospital, Toronto, Ontario, Canada.

8. UCL Great Ormond Street Institute of Child Health, London UK.

9. MRC/CSO Social & Public Health Sciences Unit, University of Glasgow, Glasgow, UK.

10. Health Data Research UK (HDR-UK), Swansea University, UK.

11. Public Health and Policy, University of Liverpool, Liverpool, UK.

12. School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester UK.

*Corresponding author: Dr Elizabeth Ford, Department of Primary Care and Public Health, Brighton and Sussex Medical School, Watson Building, Village Way, Falmer, Brighton, BN1 9PH, UK. E.m.ford@bsms.ac.uk; +44 1273 641974.

Abstract

The last six years have seen sustained investment in health data science in the UK and beyond, which should result in a data science community that is inclusive of all stakeholders, working together to use data to benefit society through the improvement of public health and wellbeing.

However, opportunities made possible through the innovative use of data are still not being fully realised, resulting in research inefficiencies and avoidable health harms. In this paper we identify the most important barriers to achieving higher productivity in health data science. We then draw on previous research, domain expertise, and theory, to outline how to go about overcoming these barriers, applying our core values of inclusivity and transparency.

We believe a step-change can be achieved through meaningful stakeholder involvement at every stage of research planning, design and execution; team-based data science; as well as harnessing novel and secure data technologies. Applying these values to health data science will safeguard a social license for health data research, and ensure transparent and secure data usage for public benefit.

Introduction: health data science as a UK national priority

Healthcare and health research are being rapidly and dramatically transformed by the increasing availability of electronic data and the extraordinary advances in computational power required to process them. New knowledge is being generated by significant advances in health informatics, in data capture and curation, knowledge representation, and data analytics. These advances are critical both to the delivery of healthcare for the population of the United Kingdom (UK, population 66.6 million) and to the digital health and life sciences sector, one of the most dominant economic sectors in the UK, estimated to be worth approximately £67bn in 2016 (1). Due to demographic and population pressures, there is a substantial need for greater efficiency within the UK health system. In the UK National Health Service (NHS) the Five Year Forward View and Personalised Health and Care Plan 2020 set out a strategy for the NHS in England to revolutionise health and care for patients through the adoption of digital tools and technologies. Specifically, NHS Digital's Data Services Platform is seen as the infrastructure for a future national Learning Health System (2).

In March 2013, the Farr Institute of Health Informatics Research - a collaboration of 21 academic institutions and health partners across the UK - was established to bring about a step-change in the harnessing of data to improve population health, address health inequalities and to drive efficient service provision. Farr's vision was encapsulated by its '#datasaveslives' tagline. The Farr sat within a growing international, interdisciplinary community of those concerned with 'population data science' (i.e. the science of data about people) (3); which in turn sits within the wider 'Big Data' landscape of those seeking to use increasingly rich digital data for a wide variety of purposes.

Since its creation, Farr researchers have successfully harnessed the power of data to study a wide range of clinical and public health issues. An important focus of the Farr has also been bringing together datasets from different healthcare and administrative sources through confidential record linkage. These datasets have included hospital and primary care data, as well as public health (e.g. screening programs) administrative and educational data. While the main focus of the Farr Institute was on quantitative data science, qualitative work provided additional insights into public attitudes to data science (4) into the underlying drivers and motivators of behaviours (5), and enthnographic insights into the data science process (6). Through these approaches, Farr researchers published 238 peer-reviewed papers during 2016-2017 alone (7).

However, there remains concern that the opportunities made possible through the innovative, efficient and secure use of data are still not being fully realised, resulting in avoidable health harms (8). Additionally, attempts to broaden the range of datasets which are used within data linkage studies have made slow progress. This includes data from novel sources such as social network platforms, commercial transaction records, smartphones and wearable devices, sensors, and internet connected devices, both in the home and the public realm. Research has demonstrated that these data can be used to inform improvements in population health, for example by characterising disease outbreaks in near real-time (9), identifying side effects of medications (10), and predicting clinical deterioration after hospital discharge (11). The UK (and England in particular) also falls behind other countries in linkage of cross-sectoral data to inform analyses of health (e.g. from employment and criminal justice systems).

In March 2017, Health Data Research UK (HDR-UK) was announced as the successor to the Farr Institute, and its scientific programmes were formally launched on 1st May 2018. The inception of HDR-UK is an opportune time for considering priorities for the future. In this paper, we, the members of the first cohort of Farr 'Future Leaders', consider what the most important priorities might be for achieving a sustainable step-change in productivity in health data science in order to enhance the health of the population.

Vision: Our Data, Our Society, Our Health

Our vision is for **a health data science community that is inclusive of all stakeholders, working together to use data to benefit society through the improvement of public health and wellbeing.** We want health and population data science to be supported by, guided by, and of direct benefit to, as much of society as possible. The tagline "**Our Data, Our Society, Our Health**" emphasises the role of society in delivering inclusive and transparent health data science.

This new social contract rests on meaningful stakeholder involvement at every stage of research planning, design and execution, requiring a broad range of researcher skills to ensure our infrastructure, governance, analysis, data management, information security and communications are all efficient, appropriate, and closely aligned to stakeholder priorities. Our vision thus needs to be communicated and executed in a manner that is clear, transparent and inclusive. We will take responsibility to ensure that the general public have the information and opportunity to understand, engage with and benefit from the outputs of science based on their data.

Current barriers to achieving our vision

We perceive a number of barriers to achieving the vision of inclusive and transparent health data science, including obstacles to data access, variability in data quality, current skills and capacity, and importantly, managing of public trust in health data research.

Data access barriers

A key issue for researchers in health data science is the risk inherent in securing access to data. Data access can be blocked by "hard" barriers such as restrictive legislative clauses or financial and technological restrictions imposed by electronic medical record (EMR) vendors (12), and "soft" barriers such as risk aversion to data sharing on the part of data custodians or ethics committees. These barriers curtail research through blocking access to data, but also through delaying access to data to an extent where it threatens the viability of typical grant-funded research or PhD projects (13).

The thinking behind these barriers may be explained by *prospect theory*, which is widely used in behavioural economics, and identifies loss aversion as a significant and consistent cognitive bias when choices are made in the face of uncertainty (14). There is an over-emphasis on perceived negative impacts of prospects that are of low probability but can involve large losses, leading to inertia. We consider that risk aversion and subsequent inertia has resulted from data owners responding to rare events such as data breaches (15) (e.g. HM Revenue and Customs ceasing to share data following the loss of CDs containing the records of 27m UK taxpayers (16)), public scandal (e.g. changes in NHS data sharing

practice following the Partridge Review (17)) and legislative change (particularly when there is limited regulatory guidance on how to interpret the new legislation). Inertia and risk aversion are also introduced where governance challenges are identified, but the data owner is either unfamiliar with technical solutions or is not certain that these are compliant with regulations. Even the introduction of legislation designed to facilitate data science, such as the UK implementation of the EU General Data Protection Regulations (GDPR) and the UK Digital Economy Act 2017, lead to uncertainty, and potential harms while regulators and data owners work to develop new codes of practice and establish new norms.

Currently, data access is hampered by a lack of, or inconsistent use of, national standards. For example, UK government departments differ in expectations as to how researchers demonstrate information security standards, with some data owners accepting ISO27001 certification (18) (arguably the leading international information security accreditation standard), others requiring NHS Information Governance Toolkit (19), and other departments requiring evidence of compliance with HM Government Security Policy Framework (20). The lack of consistency increases administrative burden on the research community and makes data application processes and timescales unpredictable.

These inconsistencies again contribute to risk aversion by decision-makers who are minded to take a cautious interpretation of legislation, are not clear what the law allows, and may lack familiarity with innovative data science proposals and technologically driven solutions. Similarly, local ethics committees may have poor knowledge of data sharing and data science projects because of the low throughput of these projects in their geographical area. Furthermore, some centralised data access bodies operate in opaque ways, which are difficult for researchers to navigate and lack opportunities for meaningful engagement.

A further inconsistency lies in differing government structures and governance environments across the four home nations comprising the UK. Using research access to hospital and registry records as an example: In England, access to hospital data is obtained through NHS Digital, with applications considered by the Independent Group Advising on the Release of Data (IGARD); In Scotland, data can be requested from the Information Services Division and the National Records for Scotland, where applications are considered by the Public Benefit and Privacy Panel for Health and Social Care (PBPP) via eDRIS; In Wales, researchers gain access to data held in the Secure Anonymised Information Linkage (SAIL) Databank, held at Swansea University, through an application to an independent Information Governance Review Panel; In Northern Ireland decisions are ultimately made by the medical director of the care organisation. There are many settings where conducting research with data from several jurisdictions is valuable and/or unavoidable (e.g. following participants who move to a different region in the UK or including participants from all regions of the UK to enable investigation of rare conditions or to help draw a representative national sample), but currently this requires multiple approvals with inconsistent mechanisms (21). In addition, ethics committees and other governance groups such as the Health Research Authority Confidentiality Advisory Group have requirements that can conflict with the data providers', e.g. in wording for privacy notices in line with GDPR requirements. It is not currently clear which (if any) decision has primacy, and researchers may have to provide several iterations of documentation to satisfy all parties. This barrier to efficient access to records, or an efficient mechanism for whole population sampling, has been recognised by the UK Economic and Social Research Council (22), who propose a national population 'spine' as part of the means to address these inconsistencies.

The upshot of these inconsistent approaches is that while there is clear support for increased use of routine data in research aiming to improve the public good at the highest levels (23), this is not manifesting itself in consistent and timely access to data. For example, a parliamentary report from the select committee on artificial intelligence (24) fears the benefits of data science to patient care could be stymied by a lack of a consistent approach to data-sharing arrangements between NHS organisations and developers of artificial intelligence.

Some new data sources, for example social media, are subject to few restrictions or guidelines as to their appropriate use. This has resulted in a culture of self-governing in which researchers make decisions about how to act ethically on a case-by-case basis. This has led to misuse and scandal (25). Indeed, many publications based on social media data do not mention ethical issues or simply state that consent was not required because the data were available in the public domain (26). However, this is a rapidly evolving space and there is increasing recognition of the need for ethical governance structures and guidelines for research based on internet-generated data (26).

Data quality barriers

To maximise the public benefit of health data research, it is imperative that data quality issues or biases within the data are understood and accounted for in the analysis process. There are challenges in using data not originally collected for research, hence there is a large body of literature focused on assessing the validity of routinely collected health data (27,28). This is also now extending to validation of some innovative data sources, such as sensor data (29). However, there is a rapidly growing number of new data sources for which the limitations and potential biases have not been identified (30,31). A particular challenge for users of many non-traditional data sources lies in creating replicable case or concept definitions, which are critical to understanding the extent to which the data source offers unbiased and complete information on the topic under study (31). Linking novel data, e.g. social media data, into richly phenotyped cohort studies, with alternate sources of exposure or outcome data, provide opportunities to test the ‘ground truth’ of assumptions made elsewhere solely using novel data (32).

The potential for research which benefits population health is increased exponentially with the linkage of data across multiple sources, but a lack of unique and accurate identifiers to link data across different sources can affect the quality of linkage (33). In turn, linkage error can undermine the representativeness of analyses, for example, by excluding data from hard-to-reach populations (34,35). In many jurisdictions, separation of linkage and analysis processes is recognised as good practice for protecting privacy (36). However, this can lead to a ‘black-box’, with researchers finding it difficult to obtain the information necessary to evaluate data quality and to provide transparent reporting that allows other researchers to reproduce and validate the research (37).

Skills, progression and team working across disciplines

Health data science is widely recognized as a difficult area to train in or move into from other data domains, due to a lack of training opportunities, investment, and structured opportunities to discipline-hop from other data intensive fields. Universities often face structural and logistical challenges that prevent them from offering effective cross-

disciplinary postgraduate courses in the area. This has resulted in a shortage of individuals with skills and expertise to innovate and maximise benefits of big data in health. This situation is exacerbated by industry demand for individuals with these skills; we recognise this is both a drain on University-based research, yet also a major driver for government investment into health data science.

The rise of the 'data scientist' as a distinct role - where expertise is centred on data and data systems, rather than aligned roles centred on statistical, epidemiological, biological or social science expertise – has proved difficult to accommodate in traditional university career pathways. This is increasingly the case as larger research initiatives (such as HDR-UK substantive sites, or large longitudinal studies) employ specialists to develop data linkages, build datasets, and build advanced systems to manage the storage, sharing and use of these data, while ensuring compliance with relevant ethico-legal requirements. While data scientists may have a traditional 'academic' pathway (PhD study, lectureship or fellowship), they may also have more of an IT or data administration role. Due to these hybrid academic/administrative functions, progression routes, and metrics of success, can be unclear; those in senior data scientist roles may not meet either traditional research metrics (PhD, publication history, grant income) or traditional university administration metrics (managing large teams or budgets).

Researchers who work in cultural silos are unlikely to maximise the potential of patient data. However, at the current time, setting up effective cross-disciplinary groups can be challenging, due to the lead-time of learning to 'speak each other's languages', understanding each party's data management and analytical approaches, and aligning aims and priorities.

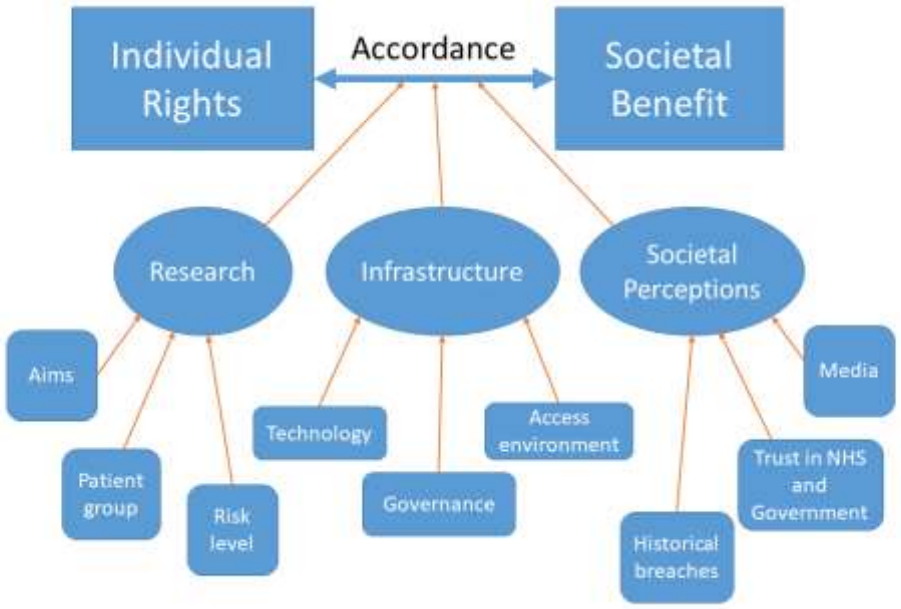
Public and patient trust

There are several important barriers in terms of gaining public trust for researchers to access health records on a large scale. Whilst there is good evidence that the public support the secondary use of routine data (4), previous highly public cases (e.g. care.data scheme to centralise primary care records in England (38), Google use of Royal Free data to develop artificial intelligence algorithms (39)) place this public support and trust at risk. Furthermore, wider misuse of data (e.g. the use of personal information from Facebook to target election campaigning (40)), and the increasing monetization of data and advice (e.g. pharmaceutical company donations to patient advocacy groups in the US (41)) risks reducing public trust for all complex data (re)use. Any loss of public trust in the NHS and universities as data custodians will impact on the ability of clinicians to deliver effective, efficient and safe care, and the ability of academics to access and use patient data for research.

Decision-makers are currently very conservative in allowing data to be shared or linked across different sites or trusts, because of perceived privacy risks. As previously discussed, this may be driven by fear of negative news headlines and the large financial and reputational costs of data breaches (42). However, there has been a notable failure to balance this against the ethics of data non-use: not sharing data may be actively harmful, and cost lives, if progress in research is not made (8). Unfortunately, this concept of harm through restrictive use of data barely weighs within the current ethics discourse, and the major focus on ethical review relates to the possible harms from privacy breaches. Decisions about data linkage and sharing for research must balance individual rights and societal

benefit, and in each case evaluate the risks and rewards of the particular research project, the data security infrastructure, and societal perceptions, as depicted in Figure 1.

Figure 1. Influences on achieving balance between individual privacy and societal benefits when considering data use.



Inclusive involvement of stakeholders

In order to deliver our vision for health data science that is inclusive of all stakeholders, we propose new approaches for engaging with the public, data owners and regulators; managing diverse and inclusive research teams; and enabling transnational health data science.

Engagement with the public and securing public trust

Two key issues have been identified in relation to trust around use of patient data: the public must trust research organisations' competence with data handling, and must trust their motivations for data analysis (4). To facilitate that trust, we must understand the level of control patients desire over their data, understand how to obtain and maintain public trust, connect our efforts to individual communities, and promote the societal benefits resulting from our work in order to secure a social licence for health data science.

Researchers can also maintain public trust and support by having a better understanding about what level of control patients wish to have over their data and how they wish to achieve this. At present our understanding of whether patients want granular control over their data is limited and there are several areas of work to be undertaken. Some evidence suggests the public are most in favour of an opt-out system for data usage after deliberative or educational events, such as citizens' juries where they have had the chance to explore the issues and understand the problems of selection bias in an opt-in model (43). This even

extends to vulnerable and marginalised groups such as people who have experienced homelessness (44). Communication and clarity of argument is key here however, as GDPR has created an expectation of opt-in mechanisms - which are now required in online and commercial contexts, but not for researchers using data for the performance of a task carried out in the public interest and for scientific research involving sensitive (e.g. all health) data (45). The data-science community will need to publicly articulate the need for opt-out approaches demonstrate the societal benefits these bring, and co-design alternate safeguards (if needed) with the public.

At present the UK public are wary of NHS data handling competence (46) and many report not trusting the motivations of private sector or government organisations (47). Conversely, and perhaps surprisingly, the public also express trust for clinical and academic institutions to appropriately use health record data, because of public benefit focused motivations (48). One report investigated public attitudes to sharing data with commercial entities, and concluded that a strong case for public benefit is the most important factor for most people to agree to research in this context. Without it, data use by any organisation is rarely acceptable (49). There is a recognised need for better communication with the public about the use of health data in private and public sector collaborations (50). It is not currently clear what the public perceive as a 'fair usage' of data for a private technology company compared to that for the NHS, a question that should be investigated further as a matter of urgency given rapid developments in this field. Options may include the primary research partnership being between the NHS and a university, and any private company wishing to access the data would need to approach this established collaboration as a third party, and show investment in the research within one or both of the public organisations.

The social landscape around data sharing is rapidly evolving and hard to predict. However, many researchers working in the field of public engagement are operationalising the theory of social licence to inform dissemination of plans for data-sharing schemes. First developed around ideas of corporate social responsibility, social licence theory proposes that the public expect that, in some circumstances, the conduct of groups or organisations should go further than the requirements of formal regulation, towards voluntary adherence to social codes of trustworthy and responsible behaviour (51). Where the public are satisfied that the motivations of the organisation are trustworthy, they grant a "social licence" to operate. Securing and maintaining public trust, and thus a social licence, for the use of patient data for research must be present in all endeavours within the health data science community (52). This approach has been pioneered at a project level by the National Institute Health Research's INVOLVE programme that aims to move researcher involvement of patient/participants from engagement, through involvement, to meaningful research co-design. Several recent initiatives show how this may be scaled to regional and national programmes. The 'Born in Bradford' cohort study have invested considerable resource in building a sense of community around their research (partly through the Connected Health Cities initiative). They are reinforcing the visible benefits to the Bradford community by moving from an observational research design to including an experimental component which aims to improve service provision to families with young children (53). The Wellcome Trust funded "Understanding Patient Data" initiative (54) has aimed to scale data use into a 'national conversation' and to improve the clarity and consistency of communications with the public regarding the use of their data and also to tap into public altruism by producing a series of videos showing how patients have shared their data for public benefit.

The public are often willing to share their data for the benefit of a “community” they belong to. Community can be defined in several ways. Firstly, as a shared geography; people feel a sense of belonging to a place. Aiming to achieve community buy-in this way works well for small countries such as Wales, or regions of England, and the Northern Cities, as the “diameter of trust” is thought to extend to 2-5 million people (55). This may, however, limit community buy-in national English or UK research initiatives, where the population is >50 million. As such, national cohort studies such as the National Survey for Health and Development, or Twins UK, have worked on fostering a community around cohort membership, rather than around a geographical area. Secondly, a shared medical interest may foster a sense of community interest. There have been several examples whereby medical research charities have achieved a critical mass among the patient community who support their data being used for research into particular conditions, for example, the Multiple Sclerosis Register (56). We would aim to encourage a sense of responsibility for community-building by researchers, with a diverse range of stakeholders, not only as a way of creating a greater sense of ownership of the research by the public, but also to better meet the needs of these groups of patients.

Health researchers must actively promote the societal benefits arising from their research, so that it becomes part of the public consciousness that research using patient data results in benefits for society. An Understanding Patient Data initiative stipulates that researchers should routinely acknowledge their sources of data in publications and press releases, with the view that by continually acknowledging the use of such data, the public will see that it is being used for the common good, and thus feel more positively about its use in general (54). To maximise public involvement, researchers should provide details of their research in open and accessible manner through public web resources and public facing engagement events (e.g. Pint of Science, Cafe Scientifique etc.). Where a sense of community has been fostered, it will be most obvious who the key stakeholders are and communication can be tailored to their needs and preferences.

Engagement with data custodians, regulators and those developing data legislation

Negotiating access to data necessitates substantial engagement with data custodians, but it is not often that researchers engage directly with regulators setting the codes of practice or with the policy makers who develop legislation. However, effective data science will need to address ‘hard’ factors, such as restrictive legislative clauses, and it is necessary to galvanise diverse groups of stakeholders, including patients and the public, to lobby for and facilitate legislative change. The Wellcome Trust took this lead in galvanising the EU data science community to press for research exemptions for onerous data protection clauses, such as the requirement for specific opt-in for all research use of personal data, in the recent EU GDPR. A further example is seen in the case of the CLOSER longitudinal cohort consortium (57) who worked with the Cabinet Office to develop clauses within the Digital Economy Act (58) to facilitate the sharing of routine records for research purposes. Within teams developing legislation, it is therefore crucial to include data scientists with an applied understanding of the barriers faced and potential solutions, and how those issues manifest within a particular research domain.

Established facilitators who enable researchers to gain data access also show promise in increasing efficiencies. Recently, the ESRC have established the Administrative Data Research Partnership to act as the facilitating organisation enabling data-scientists to access routine records collated by the Office for National Statistics. Examples of further good practice and impact in this field should be collected and made available to support future funding for this type of activity, and increasing the efficiency of the process for comparable projects. In addition, there should be incentives in place which encourage data custodians and providers to manage and share data in accord with existing standards. Such open data management incurs a financial cost, especially if security infrastructure around it is to be maintained. Thus, going forward, it is important that the expectation of data-sharing be embedded in the "data culture" among all stakeholders, and financial provision be made to enable such open data.

The inclusive data science team

Addressing interdisciplinary and cross-sectorial challenges within contemporary data science requires building teams with a diverse range of skills, and developing methods for staff retention, recognition, and progression.

Teams, whether project-based, institute-based or those working collaboratively across institutions, are likely to require a mix of applied investigative skills, clinical expertise, data management, governance and informatics expertise. This will include scientists from traditional clinical and academic backgrounds moving between medical research, informatics and other disciplines. For example, astrophysicists can have a role in transferring 'big data' skills into medical informatics (59). Social scientists have important roles, as exemplified by the CLOSER project, linking biomedical with equivalent social studies to share best practice and develop joint infrastructure, while lobbying for data access improvements to link NHS data to cohorts (57). Sociologists help us to understand how the general public relate to the use of their personal data, and will be instrumental in generating the new governance needed to realise societal benefit by optimising use of cohort, routinely collected, and other data more generally (60). Involving other disciplines widens the types of data linked to health data and increases evidence of the impact made by social or environmental systems. For example, geographers have worked on household and individual level data linkage to investigate the influence of access, or lack thereof, to health promoting or demoting facilities (61). This is particularly important in the hope of realising a shared responsibility between health care providers, patients, and local government who help shape local environments (62,63). Additional support will be needed from ethicists, security experts, project managers, public relations and communications experts, and contract lawyers, amongst others. Building such teams is challenging given the competition between selecting in-demand skill sets and managing these across the project life course.

Effective mechanisms are also needed to enable retention of skilled team members, both through direct compensation to address the challenge of wage competition with the private sector, and recognition within universities that data scientists have particular value that is often lost through the challenges of short-term contracts and restraints to career progression. Clear metrics are required to recognise contributions from those with less 'traditional' academic roles in order to build morale and support career progression. Our recommendation is that metrics are developed to establish equivalent recognition between publishing well-managed datasets to publishing journal articles. These would need including

in all key metrics reviews (e.g., we support that a ‘data resource’ publication describing a new and valuable dataset should score well within the Research Excellence Framework (REF), (64)). Assessment of research contribution at both the individual level (e.g. for promotion) and at the institutional level (e.g. in the REF), must value contributions such as setting up and managing data security infrastructure and data resources, developing data-sharing policies, and engagement with stakeholders. In time, equivalent measures to assess quality and impact would develop and recognise good practice across the data lifecycle (i.e. from the design of data collection tools through to the documentation and archiving of collected data).

For building capacity in health data science, we propose three key schemes. Firstly, a doctoral training program in health informatics and health data science is needed to generate future specialists and leaders in the field, to be deployed into academia, industry and the public sector. Secondly, a professional development program aimed at upskilling analysts and informaticians working in the NHS is needed to deliver innovation into the health system. Thirdly, postdoctoral fellowships in health data science, together with networks for early career researchers and future research leaders are needed to grow a national research community. This approach has been exemplified by The Farr Institute, and taken up further by HDR-UK.

Enhancing the impact of health data science by going transnational

Collaborating across international boundaries can contribute to enhanced productivity and impact. Our proposal is for strategic funding to support transnational research to enable meaningful and focused collaborations. Short courses and transnational visits are excellent means to promote knowledge exchange and seed collaborative projects. Since scientific collaborations are often based on the flow of researchers between institutions, it is important that UK’s immigration policy recognizes health data science as a key area for highly-skilled migrant programmes.

The inclusion of data from multiple populations increases statistical power, thereby improving the precision of estimates and allowing for the investigation of relationships that could not otherwise be examined due to rare exposures, outcomes, or both. Including data from populations with diverse genomic or social backgrounds increases the generalizability of the research findings. Additionally, transnational research provides opportunities to conduct “natural experiments”, as social and environmental contexts, health infrastructure and payment systems differ across borders (65-68).

As difficulties with sharing and linking data within national boundaries can be magnified when attempting to scale up across international borders, transnational research based on the inclusion of unit record data from the UK in a centralized repository is unlikely to be feasible in the near future. It is, however, currently feasible to rapidly conduct multinational studies by applying a common protocol or common data model across distributed networks, followed by pooling the results. There are many successful examples of multinational research arising from such distributed networks, including some involving data from regions of the UK (69-71), and a number of EU-funded transnational health data research projects (72-74).

Transparency in health data research

We consider transparency in the context of data flows, data processing and research outputs, proposing new approaches for governance structures and pathways, and ensuring openness of data management and research software used.

Transparent governance structures

Our proposal is for increased transparency of governance systems in place for personal health data, ensuring their agility in dealing with emerging technologies, and working with patient and public groups to create pressure to address legislative challenges.

For transparency and accountability, we must ensure that clear information describing data flows, data sharing agreements, research objectives, results, and their clinical impact, are not only made publicly available but actively promoted. For example, researchers must communicate the reasons why consent systems are designed in particular ways (e.g., to reduce the impact of bias) and the safeguards in place to control patient privacy (e.g., disclosure control mechanisms), the ethics process for accessing data, how research teams are trained in governance, and constraints placed on data access.

Earlier we described how ethico-governance factors act as hard or soft barriers to data science. Addressing governance issues is likely to remain challenging given that governance expectations of ‘good practice’ are dynamic, adapt to technological change and changes in social perceptions on data use. Present-day examples include the need to develop frameworks to regulate and govern the use of artificial intelligence in health-care provision and also for the use of social media, commercial transaction data and continuous monitoring data from sensors and ‘internet of things’ devices. Therefore, any given future governance model needs to be agile, assess stakeholder expectations and be responsive to emerging challenges (75). The feasibility of implementing models based on frameworks such as these is constrained by the diverging requirements imposed on researchers by the data custodian community.

It is essential to standardise access to data through adopting co-ordinated standards across research communities and co-ordinated research infrastructure (e.g., to build on the NHS’s integrated ethics application system (IRAS) as a one-stop-shop for all ethics and data applications). Flexibility should be further enhanced by separating evaluation and accreditation of data-handling for individuals and research projects. In this scenario, a research project would still need ethical review, but would be carried out only by individuals who have undergone training (e.g. in confidentiality, disclosure control), have valid contracts with bona-fide institutions and have achieved the thresholds required to hold some form of ‘data research passport’. This would facilitate collaborative team working across institutions and reduce the overheads involved with setting up individual projects and their management.

Honest data management

Today’s data science community has access to information technology and governance solutions needed to ensure data science takes place in an ‘honest’ manner: that is, the assurances provided to the public and data owners when the data was acquired and (re)used in research are transparently upheld and auditable.

Current attempts to reduce disclosure-risk often impact on the granularity of individual level data, as data are made more anonymous by aggregation or stripping out details at the patient level. We believe a better way to protect patient privacy but retain granularity is the Data Safe Haven or Trusted Research Environments (TRE) approach (7,78), where data are kept in their full resolution but access is restricted to bona fide users (75). This is illustrated by the capabilities within the SAIL databank to conduct address level data linkage, using a novel data linkage system containing *anonymised* patient address data (79,80). Data provider and societal reassurance is further secured by controls over access (e.g. researcher training, auditing, contracts, penalties, output checks). The capabilities provided by the TRE model have enabled complex intervention and natural experiment evaluations that are now providing valuable evidence to guide public policy governing environment,(81) education, and other large societal systems (81-83). Although access has been a problem in the past, with unplanned outages and restricted working hours, substantial recent investment means these platforms are more user-friendly than ever and access will likely continue to become more streamlined and more reliable.

Models for this type of socio-technical infrastructure extend beyond the technology needed to keep data safe while being used for research. They include platform elements that promote data discovery and access to comparable data harmonised across many studies (e.g. Dementias Platform UK). These focused research/data/clinical environments maximise the opportunity for data science to have rapid translational benefits. Data 'streams' should flow into the system, within which methods and models sit, and are shared between experts.

For example, Connected Health Cities is a learning health system in the north of England bringing together subject specific expertise with data and skills analysts centred on regions (76), aiming to minimise the 'data-action latency'. This means that insights gained from the data should translate rapidly into action and impact in the real world. This combination of expertise and workflow use increases the efficiency of the system and hence the use of skilled individuals. An e-lab collates data and expertise around a scientific question or domain, e.g., STELLAR is an e-lab/platform built around endotype discovery in asthma using harmonised information from cohort studies and linked routine records (77). By bringing together researchers working in a similar field, sharing their methods and data, Connected Health Cities is able to improve the scale, replicability and reproducibility of their research.

Despite these promising platforms, our ability as a community to drive methodological innovation in high quality data-linkage is sometimes limited by restrictions on who can carry out linkage. The organisations who have access to patient-identifiable data, and who act as trusted third parties for linkage (e.g., NHS Wales Informatics Service (NWIS) in Wales, NHS Digital in England) often have limited capacity for driving forward advances in linkage methodology. We recommend a higher degree of integration between researchers with data skills for developing high-quality linkage methods and these "trusted third parties" (84). We support data owners and data scientists working collaboratively (e.g., recent NHS Digital workshops on enhancing infrastructure capabilities (85)) and suggest that senior governmental infrastructure managers also become embedded in the data science community (e.g., the current NHS Digital Director of Data also has a position in HDR-UK). Where researchers have had an input to the linkage process (following appropriate regulations), opportunities have arisen for both in-depth evaluation of linkage quality, and methodological advances in linkage techniques (86,87).

Reproducible software architectures

Transparency of software tooling is increasingly recognised as a major problem in all data-intensive settings, as the increase in volume and velocity of data makes it impossible to use manual methods to track all the ways in which data is transformed and utilised. Thus, we recommend usage of standards to specify clear audit requirements from research tasks and usage of technologies such as scientific workflows, to provide an auditable trail of research data flows.

Fully transparent sharing of information between researchers requires us to also understand data trajectories after leaving data providers (37). Health data will typically be processed in a range of ways before the research analysis commences. Thus, we must adopt technological solutions to ensure information about data processing is captured, can feed the demand for descriptive information on quality and provenance of data, and encode these demands in formal guidelines. For example, the RECORD Statement makes a number of recommendations for the reporting of how outcomes and exposures are coded, the process and quality of linkage, and data pre-processing or cleaning (27). This is particularly relevant when we bring together data from different sources: methods used to format, link and manage data prior to analysis can have a large impact on results. Systematic sharing of data cleaning frameworks, data management plans, and clinical code lists pushes forward the efficiency of data science, and allows efficient validation of results in different datasets and settings (88). These meta-data may also be shared in data science repositories (e.g. Dryad (89)), where researchers are encouraged to deposit data once their work has been published.

A prospective view of recording the research workflow means capturing and documenting data processes *as they happen*. Originally developed within UK's eScience programme (90), "scientific workflows" emerged as core entities for encapsulating analytical knowledge, and have recently led to the concept of Knowledge Objects and the Common Workflow Language (CWL) standard (91). Workflows created by the company KNIME have been successfully used in sharing computable phenotype definitions as part of the eMerge project (92). Elsewhere, the CLOSER cohort consortium has developed a metadata repository using the Data Documentation Initiative (DDI) 3.1 life-cycle model (57); which is built on the concept of defining metadata as the first step of the process and then using it to drive the data lifecycle (e.g., to build online surveys, to quality check data being captured in real time, to document data files and to populate data discovery systems). Conversely, a retrospective view is dedicated to capturing, storing and analysing the audit trail produced as part of the research process. This data provenance captures causal links between algorithms, data sets and actors in a data science research process (e.g., taint analysis can be used to trace the effects of an erroneous algorithm through the system), is recognized as a key component to providing trust in the Learning Health System (93) and is supported by the W3C PROV standard (94).

Conclusions

Our vision for the future of Health Data Science in the UK is one in which inclusivity and transparency are key principles. To achieve this vision, we need to bring in key stakeholders from the earliest stages of research projects, embrace approaches that allow collaboration between providers, linkers and users, and prioritise development of research methods that

address the unique challenges in appropriate use of routinely collected clinical and emerging sources of health data. International best practice should be recognised and adopted.

Having articulated a ‘vision’ for the contemporary data science landscape within the UK and beyond, we discussed barriers restricting our vision, and proposed possible solutions. Our aim has been to use insights and experiences from the Farr Institute to help set and shape the agenda for UK Health Data Science over the course of the next few years. We suggest a new approach which prioritises **engagement with the public**, bases itself around effective and well-trained **multi-skilled teams with responsible governance** at its heart and with **sufficient data expertise** to enable the transparent and consistent capture, transformation and use of complex and diverse data.

We emphasise the need for **inclusivity** and **transparency**. Moving forward, stakeholders’ views need to be accommodated in our sector’s thinking, and ideally represented in our work. This may be either through direct inclusion in multi-disciplinary, multi-skilled teams or in meaningful involvement mechanisms, such as patient panels with active roles across the research lifecycle. The technological and governance frameworks we develop to support this work need input from all stakeholders. The operations need to be transparent in terms of data usage (to meet a governance need), data quality and provenance (to inform research analysts). Finally, research successes need to be publicised in a manner accessible to the general public, widely promoted and celebrated.

By pushing for a societal acceptance that the usage of individual’s health data in research is a vital part of the future health system, UK’s emerging Health Data Science infrastructure and stakeholder engagements will establish ‘Our Data, Our Society, Our Health’ as the new social contract to deliver a sustainable and seamless integration of clinical and research domains to improve the nation’s health and wellbeing.

Conflicts of Interest

All authors declare no conflicts of interest

Acknowledgements

This paper is the work of the first cohort of the Farr Institute’s ‘Future Leaders’ scheme. We thank Catherine Goddard, Colin McCowan, George Moulton, Paul Taylor, Athanasios Anastasiou, and Wing-Chau Tung for running the Future Leaders programme, and the support of the Farr Institute Directors in providing leadership insights. The Future Leaders programme was funded by the Farr Institute and was financially supported by the authors’ institutions or grants. All authors contributed to this paper: with all contributing to the design and framing of this position statement; with EF, AB, JB, AH and VC editing the paper; all authors contributing substantial sections of the draft, approving the final version, and taking accountability for all aspects of this work.

References

1. HM Government. Strength and Opportunity 2017: The landscape of the medical technology and biopharmaceutical sectors in the UK. The Office of Life Sciences, London, UK. [Available from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/707072/strength-and-opportunity-2017-bioscience-technology.pdf]

2. NHS Digital. Data, insights and statistics. Information and technology for better health and care. Sept 2018. [Available from https://digital.nhs.uk/binaries/content/assets/website-assets/data-and-information/dis/data-insights-and-statistics-download.pdf]

3. McGrail K JK, Akbari A, Bennett T, Boyd A, Carinci F, Cui X, Denaxas S, Dougall N, Ford D, Kirby RS. A Position Statement on Population Data Science. International Journal of Population Data Science. 2018 (Feb 22;3(1).).

4. Stockdale J, Cassell J, Ford E. "Giving something back": A systematic review and ethical enquiry of public opinions on the use of patient data for research in the United Kingdom and the Republic of Ireland. Wellcome Open Research. 2018;3:6

5. Bowes L, Evans J, Nathwani T, Birkin G, Boyd A, Holmes C, Thomas L, Jones S. Understanding progression into higher education for disadvantaged and under-represented groups. 2015 [Available from: https://dera.ioe.ac.uk/24682/1/BIS-15-462-understanding-progression-into-higher-education-final.pdf]

6. Tempini N. Science Through the "Golden Security Triangle": Information Security and Data Journeys in Data-intensive Biomedicine. [Available from: https://ore.exeter.ac.uk/repository/bitstream/handle/10871/26446/InfoSec%26Research_ICIS_2016_Final_NT.pdf?sequence=1&isAllowed=y]

7. Farr Institute. The Farr Institute of Health Informatics Research Annual Report 2016-2017. 2017.

8. Jones KH, Laurie G, Stevens L, Dobbs C, Ford DV, Lea N. The other side of the coin: Harm due to the non-use of health-related data. International Journal of Medical Informatics. 2017;97:43-51.

9. Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. Am J Trop Med Hyg 2012;86:39–45.

10. Frost J, Okun S, Vaughan T, Heywood J, Wick P. Patient-reported outcomes and a source of evidence in off-label prescribing: An analysis of data from PatientsLikeMe. J Med Interet Res 2011; 13(1); e6

11. Li D, Vaidya J, Wang M, Bush B, Lu C, Kollef M, Bailey T. Predicting clinical deterioration of outpatients using multimodal data collected by wearables. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018.

12. The College of Family Physicians of Canada. CFPC Position Statement supports access to EMR data for quality improvement and research. 2017. [Available from: https://www.cfpc.ca/position-statement-supports-access-EMR-data-quality-improvement-research/]

13. Dattani N, Hardelid P, Davey J, et al Accessing electronic administrative health data for research takes time. Archives of Disease in Childhood 2013;98:391-392

14. Kahneman D, Tversky A. Prospect theory: An analysis of decision under risk. Handbook of the fundamentals of financial decision making: Part I: World Scientific; 2013. p. 99-127.

15. Law Commission. Data sharing between public bodies: a consultation paper. 2013(214).[Available from: https://www.lawcom.gov.uk/project/data-sharing-between-public-bodies/]

16. BBC News. "Data lost by Revenue and Customs" 2007 [Available from: http://news.bbc.co.uk/1/hi/7103911.stm]

17. Partridge N. Review of data releases by the NHS Information Centre. Health and Social Care Information Centre; 2014. [Available from: https://www.gov.uk/government/publications/review-of-data-releases-made-by-the-nhs-information-centre]

18. Department for Education. The National Pupil Database and/or Linked Data. Information Security Questionnaire 2014 [Available from: https://v1.educationendowmentfoundation.org.uk/uploads/pdf/NPD_Information_Security_Questionnaire_-_with_notes_for_EEF_evaluators_-_May_2014.pdf]

19. Nwolie M, Kaliapermall V. Process for IG assurance in support of CAG and DAAG applications. 2012 [Available from: https://www.hra.nhs.uk/documents/260/cag-igt-process.pdf]

20. Cabinet Office. HMG Security Policy Framework. 2018.[Available from: https://www.gov.uk/government/publications/security-policy-framework/hmg-security-policy-framework]

21. Hardelid P, Davey J, Dattani N, Gilbert R, the Working Group of the Research and Policy Directorate of the Royal College of Paediatrics and Child Health Child Deaths Due to Injury in the Four UK Countries: A Time Trends Study from 1980 to 2010. (2013) PLoS ONE 8(7): e68323.

22. Davis-Kean P, Chambers RL, Davidson LL, Kleinert C, Ren Q, Tang S. Longitudinal Studies Strategic Review. 2017 Report to the Economic and Social Research Council. 2018. [Available from: https://esrc.ukri.org/files/news-events-and-publications/publications/longitudinal-studies-strategic-review-2017/]

23. Minister of State for the Cabinet Office, Paymaster General. Open Data White Paper - unleashing the potential. London, UK 2012 [Available from: https://data.gov.uk/sites/default/files/Open_data_White_Paper.pdf]

24. House of Lords Select Committee on Artificial Intelligence. AI in the UK: Ready, Willing and Able? London, UK 2018. [Available from: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf]

25. Samuel G & Derrick G. Social media research, “personal ethics,” and the Ethics Ecosystem [New Social Science Blog]. New Social Media. 2017 [Retrieved from http://nsmnss.blogspot.co.uk/2017/10/social-media-research-personal-thics.html. Archived at http://www.webcitation.org/6zgAWcdkQ]

26. Taylor J & Pagliari C. Mining social media data: How are research sponsors and researchers addressing the ethical challenges? Research Ethics 2018; 142(2), 1-39

27. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Medicine. 2015;12(10):e1001885.

28. Langan, S. M., Schmidt, S. A., Wing, K., Ehrenstein, V., Nicholls, S. G., Filion, K. B., . . . Benchimol, E. I. (2018). The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ, 363, k3532. doi:10.1136/bmj.k3532

29. Pires IM, Garcia NM, Pombo N, Florez-Revuelta F, Rodriguez ND. Validation techniques for sensor data in mobile health applications. Journal of Sensors Volume 2016, Article ID 2839372; http://dx.doi.org/10.1155/2016/2839372

30. van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. Journal of Clinical Epidemiology. 2012;65(2):126-31.

31. Hashimoto RE, Brodt ED, Skelly AC, Dettori JR. Administrative database studies: goldmine or goose chase? Evidence-Based Spine-Care Journal. 2014;5(2):74.

32. Framework for linking and sharing social media data for high-resolution longitudinal measurement of mental health across CLOSER cohorts. [Available from https://www.closer.ac.uk/research-fund-2/data-linkage/framework-linking-sharing-social-media-data-highresolution-longitudinal/]

33. Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., & van der Meulen, J. H. (2017). A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology. 2017 Sep 7;46(5):1699-710.

34. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data linkage: a powerful research tool with potential problems. BMC Health Services Research. 2010;10(1):346.

35. Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. BMC Medical Research Methodology. 2014;14(1):36.

36. Kelman CW, Bass AJ, Holman C. Research use of linked health data—a best practice protocol. Australian and New Zealand Journal of Public Health. 2002;26(3):251-5.

37. Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang L-C, Smith P, et al. GUILD: GUidance for Information about Linking Data sets. Journal of Public Health. 2017 *40*(1), pp.191-198.

38. van Staa T-P, Goldacre B, Buchan I, Smeeth L. Big health data: the need to earn public trust. BMJ. 2016;354.

39. Hunter P. The big health data sale: As the trade of personal health and medical data expands, it becomes necessary to improve legal frameworks for protecting patient anonymity, handling consent and ensuring the quality of data. EMBO reports. 2016:e201642917.

40. Cadwalladr C, Graham-Harrison E. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach 2018 [Available from: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election]

41. Rose SL. Patient Advocacy Organizations: Institutional Conflicts of Interest, Trust, and Trustworthiness. The Journal of Law, Medicine & Ethics : A Journal of the American Society of Law, Medicine & Ethics. 2013;41(3):680-7.

42. Information Commissioner's Office. London NHS trust fined for HIV newsletter data breach 2016 [Available from: https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2016/05/london-nhs-trust-fined-for-hiv-newsletter-data-breach/]

43. Tully MP, Bozentko K, Clement S, Hunn A, Hassan L, Norris R, et al. Investigating the Extent to Which Patients Should Control Access to Patient Records for Research: A Deliberative Process Using Citizens' Juries. Journal of Medical Internet Research. 2018;20(3).

44. Aldridge RW, Story A, Hwang SW, Nordentoft M, Luchenski SA, Hartwell G, et al. Morbidity and mortality in homeless individuals, prisoners, sex workers, and individuals with

substance use disorders in high-income countries: a systematic review and meta-analysis. The Lancet. 2018; 391: 241–50.

45. Information Commissioner's Office. Guide to the General Data Protection Regulation [Available from: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/]

46. Hill EM, Turner EL, Martin RM, Donovan JL. "Let's get the best quality research we can": public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study. BMC Medical Research Methodology. 2013;13(1):72.

47. Stevenson F, Lloyd N, Harrington L, Wallace P. Use of electronic patient records for research: views of patients and staff in general practice. Family Practice. 2012;30(2):227-32.

48. Teschke K, Marino S, Chu R, Tsui JK, Harris MA, Marion SA. Public opinions about participating in health research. Canadian Journal of Public Health/Revue Canadienne de Santé Publique. 2010:159-64.

49. Ipsos Mori, Wellcome Trust. The one-way mirror: Public attitudes to commercial access to health data. Ipsos Mori, Wellcome Trust.; 2016. [Available from: https://wellcome.ac.uk/sites/default/files/public-attitudes-to-commercial-access-to-health-data-wellcome-mar16.pdf]

50. The Information Commissioner, the Royal Free, and what we've learned 2017 [Available from: https://deepmind.com/blog/ico-royal-free/]

51. Gunningham, N, Kagan, RA and Thornton, D. "Social license and environmental protection: why businesses go beyond compliance." Law & Social Inquiry 29.2 (2004): 307-341.

52. Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care.data ran into trouble. Journal of Medical Ethics. 2015:medethics-2014-102374.

53. Dickerson J, Bird PK, McEachan RR, Pickett KE, Waiblinger D, Uphoff E, et al. Born in Bradford's Better Start: an experimental birth cohort study to evaluate the impact of early life interventions. BMC Public Health. 2016;16(1):711.

54. Understanding Patient Data. If you use patient data, acknowledge it [Available from: https://understandingpatientdata.org.uk/data-citation].

55. CIO for Health and Social Care in England. Enabling Evidence Based Continuous Improvement. The Target Architecture. 2017. [Available from: https://medconfidential.org/wp-co ntent/uploads/2017/09/2017-07-13-Target-Architecture.pdf]

56. The UK MS Register [Available from: https://www.mssociety.org.uk/research/explore-our-research/research-we-fund/search-our-research-projects/the-uk-ms-register]

57. CLOSER [Available from: www.closer.ac.uk]

58. Digital Economy Act, (2017). [Available from: http://www.legislation.gov.uk/ukpga/2017/30/contents/enacted]

59. Ford E. The ASTRODEM project aims to create a predictive model which will help general practitioners (GPs) identify patients at high risk of dementia. 2017 [Available from:https://www.bsms.ac.uk/research/primary-care-and-population-health/health-informatics/astrodem/index.aspx]

60. Murtagh, M. J., Turner, A., Minion, J. T., Fay, M., & Burton, P. R. International data sharing in practice: new technologies meet old governance. (2016) Biopreservation and Biobanking, 14(3), 231-240.

61. Fone DL, Morgan J, Fry R, Rodgers S, Orford S, Farewell D, Dunstan FD, White J, Sivarajasingam V, Trefan L, Brennan I. Change in alcohol outlet density and alcohol-related harm to population health (CHALICE): a comprehensive record-linked database study in Wales. Public Health Research. 2016;4(3).

62. The King's Fund. Shared responsibility for health: the cultural change we need. (2018) [Available from: https://www.kingsfund.org.uk/publications/shared-responsibility-health]

63. The Marmot Review (2010) Fair Society, Healthy Lives: Strategic review of health inequalities in England post 2010 [Available from: http://www.instituteofhealthequity.org/resources-reports/fair-society-healthy-lives-the-marmot-review/fair-society-healthy-lives-full-report-pdf.pdf]

64. Research Excellence Framework. [Available from: https://www.ref.ac.uk/about/]

65. Gilbert R, Fluke J, O'Donnell M, Gonzalez-Izquierdo A, Brownell M, Gulliver P, Janson S, Sidebotham P. Child maltreatment: variation in trends and policies in six developed countries. The Lancet. 2012 Feb 25;379(9817):758-72.

66. Zylbersztejn A, Gilbert R, Hjern A, Wijlaars L, Hardelid P. Child mortality in England compared with Sweden: a birth cohort study. The Lancet. 2018 May 19;391(10134):2008-18.

67. Harron K, Gilbert R, Cromwell D, Oddie S, Guttmann A, van der Meulen J. International comparison of emergency hospital use for infants: data linkage cohort study in Canada and England. BMJ Qual Saf. 2018 Jan 1;27(1):31-9.

68. Kossarova L, Keeble E. Putting rising emergency hospital admissions for children into perspective: how do international comparisons help? BMJ Quality & Safety 2017; 27:7-10

69. Charlton RA, Klungsøyr K, Neville AJ, Jordan S, Pierini A, Bos HJ, et al. Prescribing of antidiabetic medicines before, during and after pregnancy: a study in seven European regions. PloS ONE. 2016;11(5):e0155737.

70. Greiver M, Kalia S, Tobin J, Sullivan F, de Lusignan S, Cheng A, et al. People Travel, But Health Data Do Not: How we Combined and Compared PBRN EMR Data Across Three Countries. NAPCRG PBRN Conference; June 26 2018; Bethesda, NY, USA 2018.

71. Huerta C, Abbing-Karahagopian V, Requena G, Oliva B, Alvarez Y, Gardarsdottir H, et al. Exposure to benzodiazepines (anxiolytics, hypnotics and related drugs) in seven European electronic healthcare databases: a cross-national descriptive study from the PROTECT-EU Project. Pharmacoepidemiology and Drug Safety. 2016;25:56-65.

72. http://www.ehr4cr.eu/

73. http://www.emif.eu/

74. https://www.i-hd.eu/index.cfm/resources/ec-projects-results/transform/

75. Burton PR, Murtagh MJ, Boyd A, Williams JB, Dove ES, Wallace SE, et al. Data Safe Havens in health research and healthcare. Bioinformatics. 2015;31(20):3241-8.

76. Ainsworth J, Buchan I. Combining health data uses to ignite health system learning. Methods Inf Med. 2015;54(6):479-87.

77. Custovic A, Ainsworth J, Arshad H, Bishop C, Buchan I, Cullinan P, et al. The Study Team for Early Life Asthma Research (STELAR) consortium 'Asthma e-lab': team science bringing data, methods and investigators together. Thorax. 2015:thoraxjnl-2015-206781.

78. Robertson D, Giunchiglia F, Pavis S, Turra E, Bella G, Elliot E, et al. Healthcare data safe havens: towards a logical architecture and experiment automation. The Journal of Engineering. 2016;1(1).

79. Rodgers SE, Lyons RA, Dsilva R, Jones KH, Brooks CJ, Ford DV, et al. Residential Anonymous Linking Fields (RALFs): a novel information infrastructure to study the

interaction between the environment and individuals' health. Journal of Public Health (Oxford, England). 2009;31(4):582-8.

80. Rodgers SE, Demmler JC, D'Silva R, Lyons RA. Protecting health data privacy while using residence-based environment and demographic data. Health & Place. 2012;18(2):209-17.

81. Rodgers SE, Bailey R, Johnson R, Poortinga W, Smith R, Berridge D, Anderson P, Phillips C, Lannon S, Jones N, Dunstan FD. Health impact, and economic value, of meeting housing quality standards: a retrospective longitudinal data linkage study. Public Health Research. 2018;6(8).

82. Lyons RA, Ford DV, Moore L, Rodgers SE. Use of data linkage to measure the population health effect of non-health-care interventions. The Lancet 2014;383(9927):1517-9.

83. Rodgers SE, Bailey R, Johnson R, Berridge D, Poortinga W, Lannon S, et al. Emergency hospital admissions associated with a non-randomised housing intervention meeting national housing quality standards: a longitudinal data linkage study. J Epidemiol Community Health. 2018:jech-2017-210370.

84. Ray D, Roebuck C, Smith O. Delivering linked datasets to support health and care delivery and research; Health and Social Care Information Centre 2018 [Available from: https://digital.nhs.uk/binaries/content/assets/website-assets/services/dars/linked-datasets-in-nhs-digital-final.pdf].

85. New Research Innovation Workshops [Available from: https://digital.nhs.uk/news-and-events/digital-hub/new-research-innovation-workshops]

86. Harron K, Goldstein H, Wade A, Muller-Pebody B, Parslow R, Gilbert R. Linkage, evaluation and analysis of national electronic healthcare data: application to providing enhanced blood-stream infection surveillance in paediatric intensive care. PloS ONE. 2013;8(12):e85278.

87. Aldridge RW, Shaji K, Hayward AC, Abubakar I. Accuracy of probabilistic linkage using the enhanced matching system for public health and epidemiological studies. PLoS ONE. 2015;10(8):e0136179.

88. Tran DT, Havard A, Jorm LR. Data cleaning and management protocols for linked perinatal research data: a good practice example from the Smoking MUMS (Maternal Use of Medications and Safety) Study. BMC Medical Research Methodology. 2017;17(1):97.

89. https://datadryad.org/

90. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Research. 2013;41(W1):W557-W61.

91. Amstutz P, Crusoe MR, Tijanić N, Chapman B, Chilton J, Heuer M, et al. Common Workflow Language, v1. 0. 2016. [Available from: https://www.commonwl.org/]

92. https://www.knime.com/

93. Curcin V. Embedding data provenance into the Learning Health System to facilitate reproducible research. Learning Health Systems. 2017;1(2).

94. Provenance Working Group W3C. W3C-PROV. Technical Report, W3C. 2011. [Available from:https://www.w3.org/TR/2013/WD-prov-overview-20130312/]