Use of public datasets in the examination of multimorbidity: Opportunities and challenges

Christopher Boulton[1], J Mark Wilkinson[1,2,3]

[1]National Joint Registry, Dawson House, 5 Jewry Street, London, EC3N 2EX; [2]University of Sheffield, Metabolic Bone Unit, Sorby Wing, Northern General Hospital, Sheffield, S5 7AU; [3]Healthy Lifespan Institute, University of Sheffield, Beech Hill Road, Sheffield, S10 2RX

Correspondence to: j.m.wilkinson@sheffield.ac.uk

Highlights:

- The availability of large public datasets for research interrogation presents great opportunities to the multimorbidity community
- The advent of Health Data Science and machine learning as mature disciplines provide the mechanics to ask novel questions of deterministically or probabilistically linked data collections
- However, the complexity and depth of data about individuals often means that assuring anonymity can be a challenge
- There is a large body of relevant legislation and information governance requirements of which data processers need to be aware
- Similarly, data curation of generic datasets is an important consideration in the development of valid and meaningful interpretation of the wealth of available data

Summary

The interrogation of established, large-scale datasets presents great opportunities in health data science for the linkage and mining of potentially disparate resources to create new knowledge in a fast and cost-efficient manner. The number of datasets that can be queried in the field of multimorbidity is vast, ranging from national administrative and audit datasets, large clinical, technical and biological cohorts, through to more bespoke data collections made available by individual organisations and laboratories. However, with these opportunities also come technical and regulatory challenges that require an informed approach. In this review, we outline the potential benefits of using previously collected data as a vehicle for research activity. We illustrate the added value of combining potentially disparate datasets to find answers to novel questions in the field. We focus on the legal, governance and logistical considerations required to hold and analyse data acquired from disparate sources and outline some of the solutions to these challenges. We discuss the infrastructure resources required and the essential considerations in data curation and informatics management, and briefly discuss some of the analysis approaches currently used.

Introduction

Multimorbidity affects approximately two-thirds of people aged over 60 years (Fortin et al., 2012; Mokraoui et al., 2016), with complex interactions between education, behaviour, socio-economic deprivation, and frailty contributing to disease risk (Marengoni et al., 2011; Vetrano et al., 2019). Interactions between diseases may be tracked at the individual and population levels, with the metabolic syndrome and osteoarthritis as common exemplars (Anderson and Felson, 1988; Saklayen, 2018). The advent of data science (the methods of recording, storing, and analysing data to effectively extract useful information), commensurate with our capacity to manage data at scale that has increased exponentially since 2000 (Hilbert and Lopez, 2011), has led to an explosion in the discovery of such disease interactions in recent years using data mining, deep learning and big data technologies (Galetsi and Katsaliaki, 2020).

Here, the term "big data" is used to describe the study and applications of datasets that are too complex for traditional data-processing application software to adequately deal with (https://www.sas.com/en_gb/insights/big-data/what-is-big-data.html). The data may be described as "big" because of its volume, velocity (rate of accrual) or variety of formats (Figure 1). The development of this field provides a great opportunity in the multimorbidity domain to gain insights for patient benefit through the linkage and efficient analysis of large and complementary datasets (Burstein et al., 2019; Pawar et al., 2020). In this review, we consider the types of data that are commonly recorded, with some examples of their application in the multimorbidity setting. We describe the governance frameworks that control access to, and terms of use for large datasets. Our reference point is legislation that is applicable in the European Union and United Kingdom, but similar regulatory arrangements apply in most other countries. We also and outline the infrastructure requirements for managing these data and describe briefly some of the computational approaches used for their analysis.

Why study pre-existing datasets?

Reutilisation of previously collected data provides a highly cost-effective use of resources. Large datasets may provide national coverage for a particular disease area, and thus a broad geographic picture of disease incidence or prevalence and of the association between input variables and endpoints. Routinely collected data, whether for administrative or other purposes, provides real-world information at scale, and thus can provide insights into a given disease that are generalisable to other populations of similar structure. Because participation is typically passive, in that the data are often collected without the requirement for individual consent, such large-scale datasets are, by their nature, inclusive and heterogeneous.

Despite the advantage of scale, routinely collected data is also typically observational in nature. The data are not collected to answer a specific hypothesis, and may suffer from various sources of bias depending upon the population structure and the planned analysis application. The results of analysis of such datasets are inferential by association rather than causation, and thus hypothesis-generating in nature rather than hypothesis-solving. Various approaches may be taken to minimise the biases associated with routinely collected datasets or to infer causation, such as propensity score matching, Mendelian randomisation (a natural experiment design), or advanced regression techniques. A detailed consideration of their relative merits is, however, beyond the scope of this review. In contrast, data from clinical trials of an intervention used to demonstrate causation may also be made publically-available for secondary analysis. These data may be free from bias in respect of the intervention from which they were generated, but can lack generalisability because of the trial's particular design and inclusion and exclusion criteria (Jaeschke and Guyatt, 1989). Clinical trials are also typically of fairly short duration or rely on short term surrogate markers rather than real clinical endpoints, and use an intervention to which the participants are prepared to be randomised.

What types of dataset are available?

The types of data sources that may be integrated into modern analytical frameworks is limited mainly by data readability (format) and computational capacity. In the healthcare domain, data may be recorded digitally for several purposes. At the local level, electronic health records provide individual patient data for the purposes of direct patient care. Summary data of care episodes may be collected for administrative purposes, such as for billing or for resource management. These data may also be collated at regional or national level for administrative purposes and to audit patterns of disease or a healthcare system.

Datasets may be classified by the sector from which they are generated. Population registers, such as the National Population Database in the UK, record geographic and demographic data including birth, death, education, and occupational data. A list of UK government accessible datasets can be found at https://ckan.publishing.service.gov.uk/dataset. Health registers may be general (for example, the Clinical Practice Research Dataset), disease-specific (for example stroke, cancer or ischaemic heart disease), or intervention-specific (for example, joint replacement registers). Social care datasets may be identified through query to local, regional or national governmental organisations. In the United Kingdom, a national repository of health and care datasets can be accessed through NHS-Digital (https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets), the Healthcare Quality Improvement Partnership (HQIP https://www.hqip.org.uk/a-z-of-nca/) or local government through the relevant application process. In these types of dataset, the available resource is generic in nature, intended for a broad range of possible applications. Data may also be collated specifically for research purposes. Research datasets may include clinical information (patient characteristics), biological information (such as assay results and genomics data), or technical information and meta-data (for example imaging and other complex physical datasets). An example of a broad-purpose research data collection at scale is the UK Biobank https://www.ukbiobank.ac.uk/. Examples of more focussed research collections aimed at specific disease areas include the Osteoarthritis Initiative (https://oai.epi-ucsf.org/datarelease/) and the Framingham Heart Study (https://framinghamheartstudy.org/) in the United States.

What can public datasets tell us about multimorbidity?

There are many examples of the use of linked datasets to better understand the relationships between different risk factors and disease susceptibility, patient care and outcomes. For example, Wolff et al (Wolff et al., 2002) used Medicare and death data to demonstrate that in 1999, 82% of beneficiaries had 1 or more chronic condition and 65% had multiple chronic conditions. They also demonstrated the exponential increase in hospitalisations and cost to society associated with increasing numbers of co-existent conditions ($211 for those with no chronic conditions versus $13,973 for those with 4 or more). In primary care, Barnett et al (Barnett et al., 2012) used a national primary care clinical informatics dataset to show that the onset of multimorbidity occurs earlier in people living in the most deprived areas versus those most affluent, with socioeconomic deprivation particularly associated with multimorbidity that included mental health disorders. The presence of a mental health disorder increased as the number of physical morbidities increased, and was much greater in more deprived than in less deprived people.

In the case of specific disease components within multimorbidity and care, Morris et al (Morris et al., 2019) recently linked Hospital Episode Statistics (HES) and data from the Sentinel Stroke National Audit Programme (SSNAP) data to examine the effect of service centralisation in stroke treatment in Greater Manchester and found that hub and spoke models of care were effective at improving patient outcomes. Likewise, Hall et al (Hall et al., 2016) examined national cardiac data from the Myocardial Ischaemia National Audit Project to profile the characteristics associated with hospital use of primary percutaneous coronary intervention, identifying inequities in provision, with older and sicker patients less likely to receive this treatment and inter-provider variation that was not explained by confounding factors. In musculoskeletal disease, Smith et al (Smith et al., 2017) used National Joint Registry (NJR) data linked to HES to identify differences in the rate of joint replacement in England by ethnic group and found these were lower in black and Asian people versus white people, highlighting a need for further work to examine whether these differences were driven by clinical factors or by cultural or access provision between the groups. On a similar theme, Bhimjijani et al used the National Hip Fracture Database (NHFD) and HES to demonstrate the persisting effect of social deprivation on fragility fractures of the hip between 2001 and 2015. Metcalfe et al (Metcalfe et al., 2019) linked the NJR, NHFD and HES to examine differences in outcome between patients undergoing hemiarthroplasty surgery versus those having a total hip replacement after neck of femur fracture, finding a higher rate of dislocation and a lower rate of revision in the group having total hip surgery. There is also published evidence of the effect of linking national audit data with other datasets to examine nursing care (Johansen et al., 2017a), anaesthesia (Johansen et al., 2017b), geriatric medicine (Neuburger et al., 2017) and surgery (Aitken et al., 2020) (Boyd-Carson et al., 2020)).

What are the legal and governance requirements for data access?

Accessing routinely collected data for secondary purposes typically requires a series of permissions to be put into place through data request to the scientific leads of the dataset manager, the data controller (who may be different to the manager) and the controllers of any data sources that the dataset will be linked with. Where personal data are to be processed, the legal bases under common and statutory law need to be established. This may involve patient consent or application to a national government authority, such as the Confidentiality Advisory Group in England (https://www.hra.nhs.uk/approvalsamendments/what-approvals-do-i-need/confidentiality-advisory-group/) for a consent waiver. Researchers may require access to data that cannot be fully anonymised, either because a specific level of granularity is required to support the intended analysis or because they intend to link the data at a record level with other health datasets.

Much of the legal framework that governs the access to clinical datasets relates to the processing of personal data (Figure 2). For the most part, national data collection will capture data under the relevant data protection legislation. In the European Union this comprises the General Data Protection Regulation (GDPR, https://gdpr-info.eu/) ordinarily using the legal basis of a task in the public interest (Section 6.1.e) and reason of public interest in the area of public health (Section 9.2.i). Secondary users of the data may also use this legal basis if they can establish a relationship with a public authority, but may also rely on legitimate interests as a lawful basis under the law.

Fully anonymous, aggregate data is out of scope of the GDPR. Anonymous data, as defined by the Data Protection Act 2018 as "information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable" (http://www.legislation.gov.uk/ukpga/2018/12/contents/enacted). Where the link between the datasets of interest is already established and the dataset appropriately anonymised, access for third party researchers does not usually required ethics approval or a consent waiver.

Alongside the legal basis that governs the original collection, an application for using the data for a secondary purpose must also consider the legal basis that their own project operates under and the necessary data flows (Figure 3). For example, does the consent model of the original data collection permit and secondary use? And if it does, would this include a flow of identifiable data to allow linkage, or just pseudonymised or anonymised data? In some cases, secondary users need to establish a de novo legal basis under common law, either by re-consenting the subjects or by making a project-specific application to the relevant regulatory authority (CAG in the UK) for exemption under the regulations. Where necessary to preserve anonymity from the research team, a trusted third party may be used to establish the links between the datasets to enable analysis to be performed.

Alongside information governance considerations, access to controlled datasets require consideration of data security. Whilst robust organisational policies or procedures will underpin a credible approach to data security, data controllers require additional assurances by way of accreditations. In England, the national Data Security and Protection Toolkit serves as an auditable self-assessment of an organisation's information security compliance, although use of this is typically by public sector organisations. Other formal accreditations include CyberEssentials (https://www.ncsc.gov.uk/cyberessentials/overview) and ISO27001 (https://www.iso.org/isoiec-27001-information-security.html).

The complexity of these requirements should not be underestimated and success at accessing datasets requires careful planning and expert advice from information governance professionals. Data controllers are increasingly trying to reduce this burden by use of software solutions such as DataShield (Wallace et al., 2014) or OpenPseudonymiser (https://www.openpseudonymiser.org/) that allow the linkage of datasets without the transfer of identifiers. Similar solutions have been used to great effect in the SAIL database (Jones et al., 2014) in Wales. It is also sensible to explore solutions where data is accessed by applicants on a secure portal, removing the need to release datasets directly to applicants and thus reducing the risk of breach and harmonising information security declaration requirements to the regulator.

Infrastructure requirements

The analysis of large and complex datasets requires a suitable supporting hard and soft infrastructure. The hard infrastructure includes the technologies necessary for large-scale data analysis e.g. data storage, processing power and capacity, software systems and tools for analysis. The soft infrastructure includes the technical services to support the hard infrastructure, and the skills (research domain knowledge and analytical) needed to undertake valid and meaningful analysis.

High Performance Computing facilities within a trusted research environment (TRE) are essential for the storage and analysis of large-scale datasets. A TRE is certified data storage environment that provides strict security controls to prevent unauthorised access and misuse of data. This infrastructure may take the form of standalone computing facilities within the institution or a shared data and analytics infrastructure within a collaborating partnership. An example of the latter is The Greater Manchester Datawell (https://www.connectedhealthcities.org/2016/09/what-is-the-datawell/), a health and care repository that captures the city's population to enable data sharing for research.

In recent years, cloud computing has become an increasingly used efficient and cost-effective approach to the delivery of the hardware infrastructure requirements. Cloud computing refers to the provision of remote computing services – storage, processing and network systems, via an internet access point. These services can be public or organisation specific and reduce the requirement for computationally and resource intensive server capacity to be retained by research teams or data controllers. Typically, these solutions are also adaptive, making additional server space or processor capacity available on demand. Commercially available cloud services are increasing being used for access, storage and operation of big datasets. When adopting cloud based solutions, it remains important that the security and legal infrastructure of the physical servers are assured.

A third party approach data access, processing and analysis is to use a remote computing solution that does not require any investment in hard infrastructure. In this model the researcher achieves access to the data through a remote device with appropriate security controls. The NJR Data Access Portal (NJR DAP https://portal.njrdap.njrcentre.org.uk/) is an example of such a remote data access and analysis portal. The NJR DAP is hosted on a virtualised environment within a secure data centre. Access to the system is via a virtualised private network using remote desktop tools that provide the necessary encrypted communications between the system and the remote user and, through configuration of the remote desktop tools, ensure that the data and the system are protected from misuse (e.g. preventing the unauthorised upload or download of files by users). The physical infrastructure is completely isolated behind a firewall that only permits access to the remote desktop. Figure 4 illustrates the physical components making up the NJR Research Environment. Physical separation of the service architecture in this way increases the level of security, separating NJR data from the outside world.

Although the appropriate technical infrastructure is crucial to support access to an analysis of datasets, it is also important that the contractual infrastructure under which the technology is developed is robust, incorporates version control, and facilitates time-registered audit of the data access and analysis. These processes allow data controllers and analytical teams to be assured of the security of the data and the availability of sufficient computational capacity to reliably run analytical routines. Where datasets are released directly to an analytical team, these responsibilities rest entirely with the receiving organisation and careful consideration of the security setup and support arrangements is required. Remote solutions via a data access portal provides these arrangements on behalf of the analytical team, but there is still be a requirement for reliable, controlled access to the secure environment.

The tools available for analysis will vary according to the preferences of the analysis teams, but software including STATA, R and Python are all widely used. These tools vary in sophistication and will require users to be familiar with the code or structure required for the input. Where an analysis platform is provided remotely by the dataset controller, the choice of tools available may be limited due to licensing considerations. For example, in the NJR DAP, the research ready data that the infrastructure hosts is created annually and the software included for use by users comprises Office, R, Python, and Stata. However, safeguards are required to ensure that any data exported from the portal does not include source data. This can be easily achieved using a quarantine approach for automated or manual inspection before download permission request is granted.

Data curation and analysis

When data is collected at scale, it is likely that there will be inconsistencies in the recording of data, missing fields and records and general untidiness. Datasets may also be revised and evolved over multiple iterations with subtle differences emerging in the resultant data. It is important that those wishing to analyse these datasets are sufficiently familiar with the underlying dataset design to be able to model for these inconsistencies. Equally important is inclusion in analytical teams of specialist clinicians in order to give a front line perspective on how data reporting rules may have been interpreted by clinical teams at the point of collection. Data curation describes the process of turning independently created data sources (structured and semi-structured data) into unified datasets ready for analysis and for archiving.

Data quality is a key consideration in determining the utility of a dataset for a given analytical purpose or for linkage with other datasets. These attributes are the data accuracy, coverage, and completeness. Accuracy is a measure of the correctness of the data within each data cell. Coverage describes the proportion of the total population that is captured by the dataset. Completeness refers to the rate at which individual data points are missing from the dataset. Systematic biases in any of these three domains will affect the validity of inferences made from the data. For example, if data is missing from a defined geographic region or due to a batch effect, then analysis results may not be generalisable to that section of the population, and if data on a particular attribute is missing in a non-random manner, then this will bias the results of dependent analyses.

Data format. Data may be made available as summary statistics that describe aggregated information about the characteristics of a group of individuals for a particular attribute of interest. An example of this may be population mean age and a measure of the distribution, or sex distribution within a given population. This type of data, by its nature, is anonymised (provided the sample size is sufficiently large) but has limited utility for examining interactions between variables and their effects. Data may also be made available at the individual level. In this case, data-points are mapped to the individual person, allowing relationships between variables to be explored at depth.

Data pre-processing includes: 1) Cleaning (dealing with missing values, data outliers, and resolving duplication or inconsistencies); 2) Integration (combining of multiple databases or individual level files; 3) Data transformation to a common format (normalisation and transformation) so that it is ready for analysis. The process requires that all the data are formatted to a common library of language rules and definitions to facilitate the analysis. This pre-processing can be challenging when applied to routinely collected clinical records, including missing or inconsistently recorded information (Peek and Rodrigues, 2018).

Data linkage is the process of linking data from different sources. The utility of a dataset may be characterised by its likability to other datasets to create a new layer of unique information. Data linkage can substantially augment the value of participating datasets, not simply by providing additive information, but by opening up new avenues and scale of investigation. However, if done incorrectly can result in bias in the linked data (Harron et al., 2017).

Data linkage may be undertaken using a deterministic or probabilistic approach. In the deterministic approach, the datasets are linked by set of unique identifiers that are common across the datasets being linked. In this instance, subject A in dataset 1 is only linked with the other dataset(s) if there is a full match of the chosen identifiers (for example full name, date of birth, sex, address, and social security number) across the datasets. The alternative approach to linking is probabilistic (or fuzzy) linkage. In this approach a wider range of potential identifiers is used, computing weights for each identifier based on its estimated ability to identify a correct match for a given individual across the datasets. The threshold at which a record matches between the datasets is thus a function of statistical probability for which a threshold is defined, according to the rigour required by the analyst. Deterministic linkage thus follows a pre-defined rules based approach whilst probabilistic linkage allows greater flexibility for automation and speed that is balanced against the requirement for linkage accuracy. The speed and accuracy of probabilistic linkage has been enhanced in recent years by the application of machine learning approaches including neural networks and natural language processing (Tucker et al., 2019).

Data analysis. A detailed review of the analytical approaches used in health data science is outside of the scope of this article. However, expertise in big data methodologies is required to successfully link, process and analyse health informatics datasets of the scale outlined above. These skills include, but are not limited to: Management of different scale within the data (high volume low density / low volume high density; versatility with a variety of computational skills including machine learning, bioinformatics, natural language processing and computational linguistics; as well as more traditional statistical analysis techniques.

Summary and conclusions

Here we have given an overview of the opportunities and challenges that are faced when using and combining large public datasets to answer questions in multimorbidity. We have focussed on the legislation and resources required as they apply to the United Kingdom and the European Union, although similar principles apply elsewhere. We have highlighted some of the available opportunities, focussing on public datasets, where the challenges of volume, velocity and variety are more pronounced than in technically or biologically curated datasets that are prepared with research as their primary purpose. The publications referred to in the field are not intended to present exemplars, but are illustrative of the variety of opportunity available to the field through embracing big data.

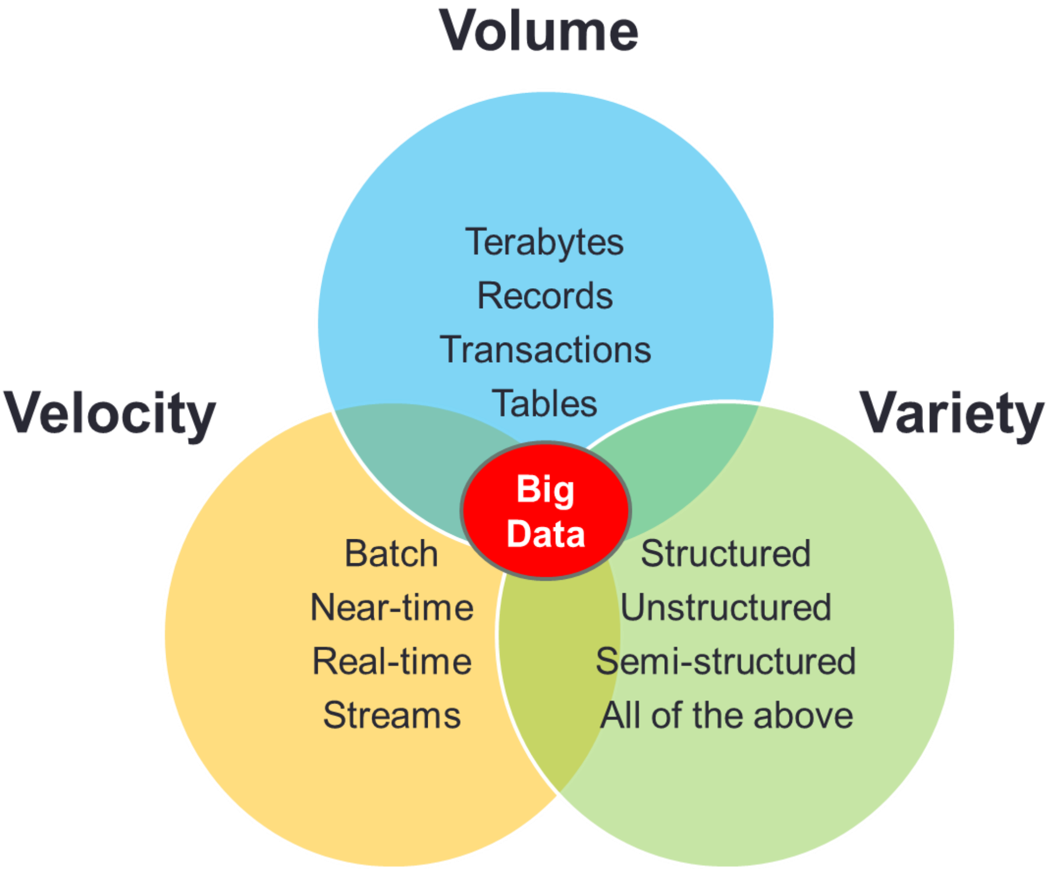
Legend to figures

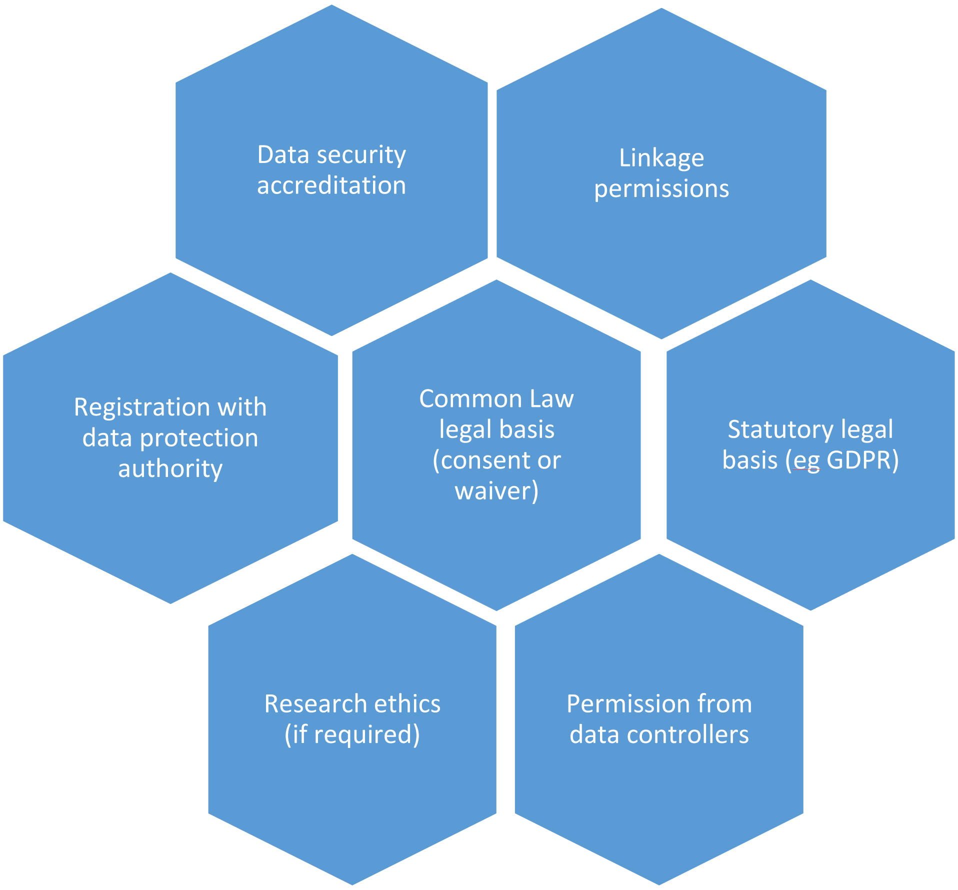
Figure 1. What is “Big Data”?

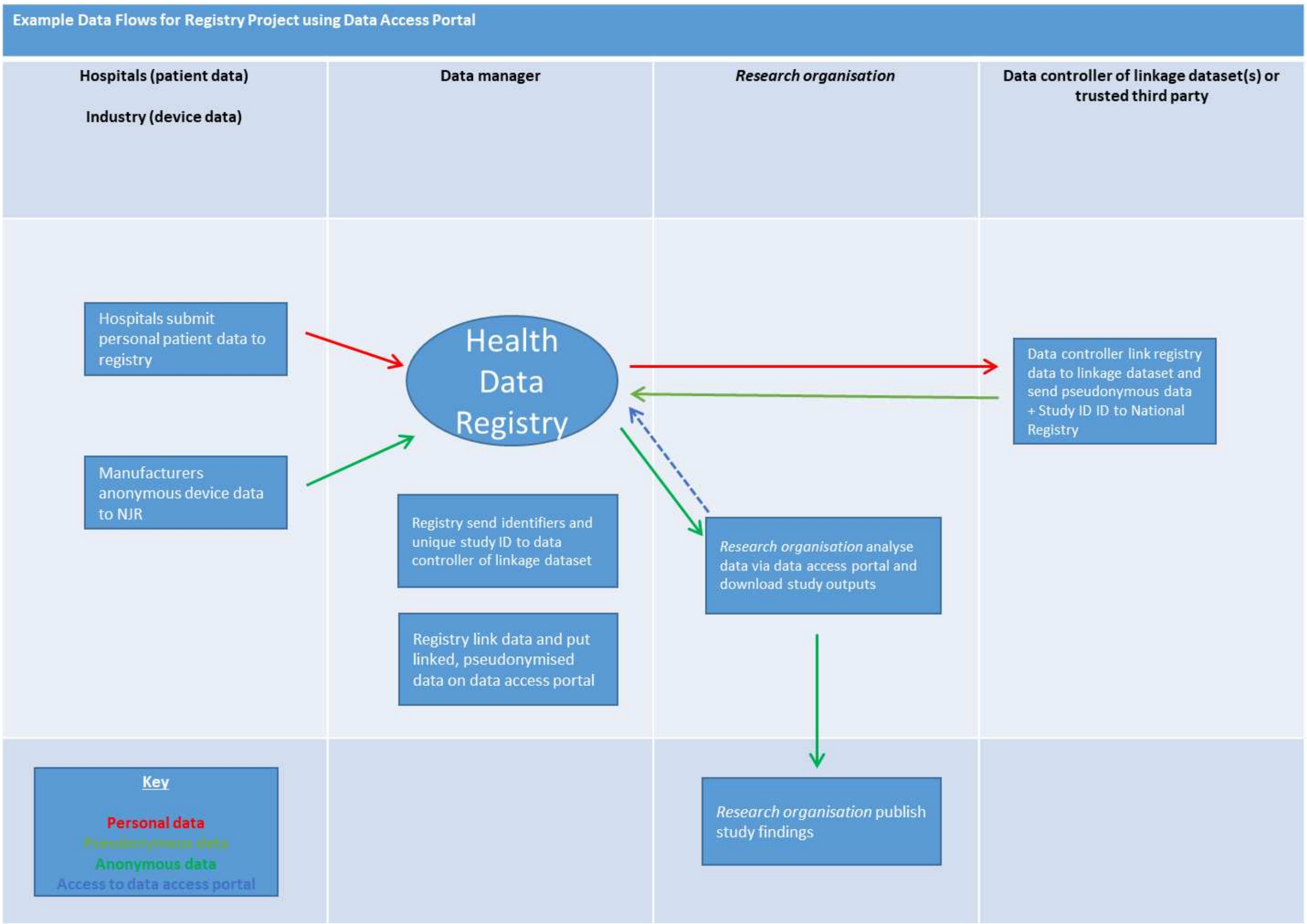
Figure 2. Key legislative areas that govern access to public datasets requiring linkage within the United Kingdom and European Union

Figure 3. Example data flows for linkage project using multiple datasets

Figure 4. Schematic representation showing the key elements of the NJR Data Access Portal, an example of a remote trusted research environment


Volume
Terabytes
Records
Transactions
Tables
Velocity
Variety
Big Data
Batch
Near-time
Real-time
Streams
Structured
Unstructured
Semi-structured
All of the above


Data security accreditation
Linkage permissions
Registration with data protection authority
Common Law legal basis (consent or waiver)
Statutory legal basis (eg GDPR)
Research ethics (if required)
Permission from data controllers



Example Data Flows for Registry Project using Data Access Portal

Hospitals (patient data)

Industry (device data)

Data manager

Research organisation

Data controller of linkage dataset(s) or trusted third party

Hospitals submit personal patient data to registry

Manufacturers anonymous device data to NJR

Health Data Registry

Registry send identifiers and unique study ID to data controller of linkage dataset

Registry link data and put linked, pseudonymised data on data access portal

Research organisation analyse data via data access portal and download study outputs

Research organisation publish study findings

Data controller link registry data to linkage dataset and send pseudonymous data + Study ID ID to National Registry

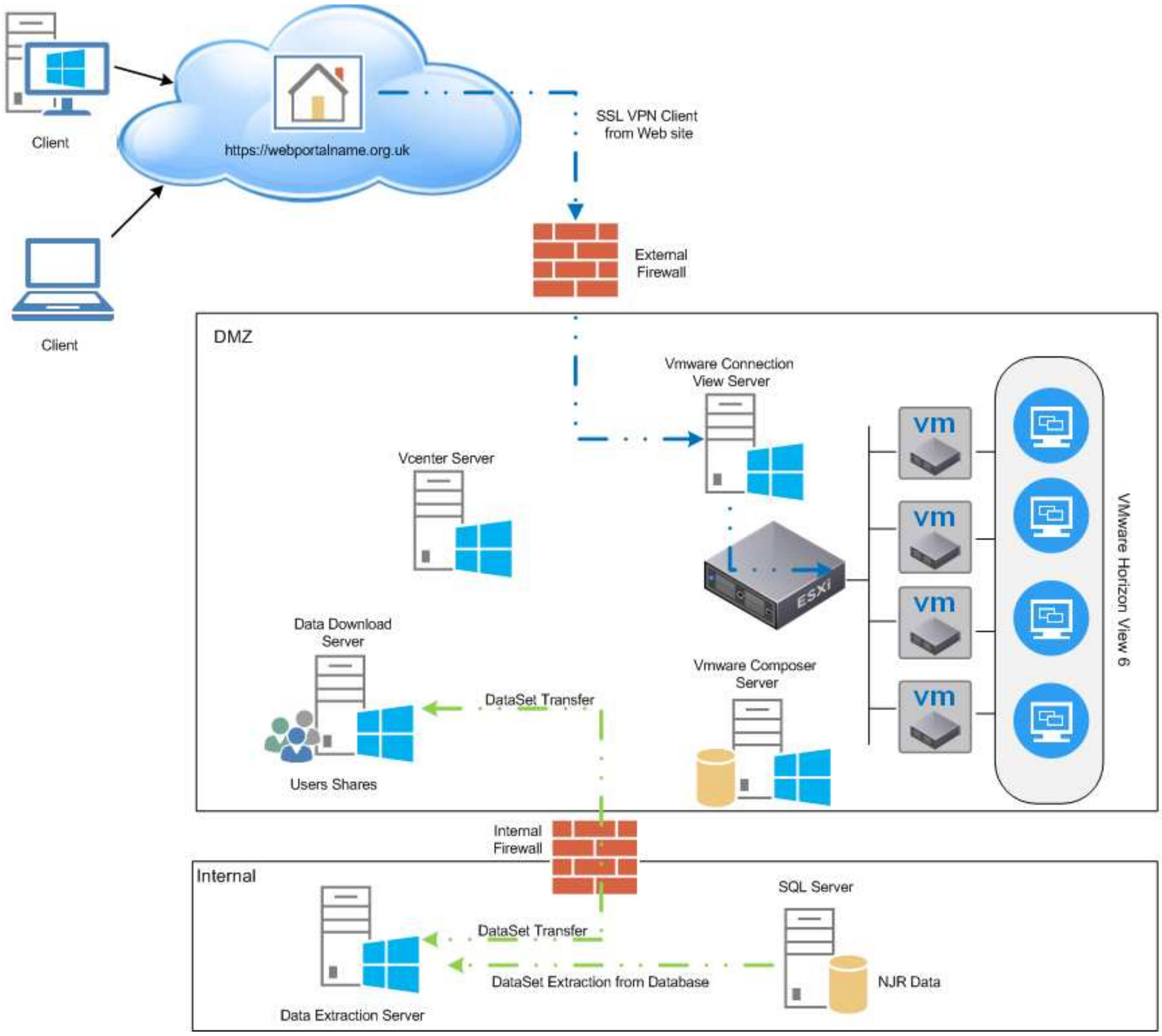
Key

Personal data

Pseudonymous data

Anonymous data

Access to data access portal


Client
https://webportalname.org.uk
SSL VPN Client
from Web site
Client
External
Firewall
DMZ
Vmware Connection
View Server
Vcenter Server
vm
vm
vm
vm
ESXi
VMware Horizon View 6
Data Download
Server
Vmware Composer
Server
DataSet Transfer
Users Shares
Internal
Firewall
Internal
SQL Server
DataSet Transfer
DataSet Extraction from Database
NJR Data
Data Extraction Server

References

Aitken, R.M., Partridge, J.S.L., Oliver, C.M., Murray, D., Hare, S., Lockwood, S., Beckley-Hoelscher, N., Dhesi, J.K., 2020. Older patients undergoing emergency laparotomy: observations from the National Emergency Laparotomy Audit (NELA) years 1-4. Age Ageing.

Anderson, J.J., Felson, D.T., 1988. Factors associated with osteoarthritis of the knee in the first national Health and Nutrition Examination Survey (HANES I). Evidence for an association with overweight, race, and physical demands of work. Am J Epidemiol 128, 179-189.

Barnett, K., Mercer, S.W., Norbury, M., Watt, G., Wyke, S., Guthrie, B., 2012. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet 380, 37-43.

Boyd-Carson, H., Shah, A., Sugavanam, A., Reid, J., Stanworth, S.J., Oliver, C.M., 2020. The association of pre-operative anaemia with morbidity and mortality after emergency laparotomy. Anaesthesia 75, 904-912.

Burstein, R., Henry, N.J., Collison, M.L., Marczak, L.B., Sligar, A., Watson, S., Marquez, N., Abbasalizad-Farhangi, M., Abbasi, M., Abd-Allah, F., Abdoli, A., Abdollahi, M., Abdollahpour, I., Abdulkader, R.S., Abrigo, M.R.M., Acharya, D., Adebayo, O.M., Adekanmbi, V., Adham, D., Afshari, M., Aghaali, M., Ahmadi, K., Ahmadi, M., Ahmadpour, E., Ahmed, R., Akal, C.G., Akinyemi, J.O., Alahdab, F., Alam, N., Alamene, G.M., Alene, K.A., Alijanzadeh, M., Alinia, C., Alipour, V., Aljunid, S.M., Almalki, M.J., Al-Mekhlafi, H.M., Altirkawi, K., Alvis-Guzman, N., Amegah, A.K., Amini, S., Amit, A.M.L., Anbari, Z., Androudi, S., Anjomshoa, M., Ansari, F., Antonio, C.A.T., Arabloo, J., Arefi, Z., Aremu, O., Armoon, B., Arora, A., Artaman, A., Asadi, A., Asadi-Aliabadi, M., Ashraf-Ganjouei, A., Assadi, R., Ataeinia, B., Atre, S.R., Quintanilla, B.P.A., Ayanore, M.A., Azari, S., Babaee, E., Babazadeh, A., Badawi, A., Bagheri, S., Bagherzadeh, M., Baheiraei, N., Balouchi, A., Barac, A., Bassat, Q., Baune, B.T., Bayati, M., Bedi, N., Beghi, E., Behzadifar, M., Behzadifar, M., Belay, Y.B., Bell, B., Bell, M.L., Berbada, D.A., Bernstein, R.S., Bhattacharjee, N.V., Bhattarai, S., Bhutta, Z.A., Bijani, A., Bohlouli, S., Breitborde, N.J.K., Britton, G., Browne, A.J., Nagaraja, S.B., Busse, R., Butt, Z.A., Car, J., Cardenas, R., Castaneda-Orjuela, C.A., Cerin, E., Chanie, W.F., Chatterjee, P., Chu, D.T., Cooper, C., Costa, V.M., Dalal, K., Dandona, L., Dandona, R., Daoud, F., Daryani, A., Das Gupta, R., Davis, I., Davis Weaver, N., Davitoiu, D.V., De Neve, J.W., Demeke, F.M., Demoz, G.T., Deribe, K., Desai, R., Deshpande, A., Desyibelew, H.D., Dey, S., Dharmaratne, S.D., Dhimal, M., Diaz, D., Doshmangir, L., Duraes, A.R., Dwyer-Lindgren, L., Earl, L., Ebrahimi, R., Ebrahimpour, S., Effiong, A., Eftekhari, A., Ehsani-Chimeh, E., El Sayed, I., El Sayed Zaki, M., El Tantawi, M., El-Khatib, Z., Emamian, M.H., Enany, S., Eskandarieh, S., Eyawo, O., Ezalarab, M., Faramarzi, M., Fareed, M., Faridnia, R., Faro, A., Fazaeli, A.A., Fazlzadeh, M., Fentahun, N., Fereshtehnejad, S.M., Fernandes, J.C., Filip, I., Fischer, F., Foigt, N.A., Foroutan, M., Francis, J.M., Fukumoto, T., Fullman, N., Gallus, S., Gebre, D.G., Gebrehiwot, T.T., Gebremeskel, G.G., Gessner, B.D., Geta, B., Gething, P.W., Ghadimi, R., Ghadiri, K., Ghajarzadeh, M., Ghashghaee, A., Gill, P.S., Gill, T.K., Golding, N., Gomes, N.G.M., Gona, P.N., Gopalani, S.V., Gorini, G., Goulart, B.N.G., Graetz, N., Greaves, F., Green, M.S., Guo, Y., Haj-Mirzaian, A., Haj-Mirzaian, A., Hall, B.J., Hamidi, S., Haririan, H., Haro, J.M., Hasankhani, M., Hasanpoor, E., Hasanzadeh, A., Hassankhani, H., Hassen, H.Y., Hegazy, M.I., Hendrie, D., Heydarpour, F., Hird, T.R., Hoang, C.L., Hollerich, G., Rad, E.H., Hoseini-Ghahfarokhi, M., Hossain, N., Hosseini, M., Hosseinzadeh, M., Hostiuc, M., Hostiuc, S., Househ, M., Hsairi, M., Ilesanmi, O.S., Imani-Nasab, M.H., Iqbal, U., Irvani, S.S.N., Islam, N., Islam, S.M.S., Jurisson, M., Balalami, N.J., Jalali, A., Javidnia, J., Jayatilleke, A.U., Jenabi, E., Ji, J.S., Jobanputra, Y.B., Johnson, K., Jonas, J.B., Shushtari, Z.J., Jozwiak, J.J., Kabir, A., Kahsay, A., Kalani, H., Kalhor, R., Karami, M., Karki, S., Kasaeian, A., Kassebaum, N.J., Keiyoro, P.N., Kemp, G.R., Khabiri, R., Khader, Y.S., Khafaie, M.A., Khan, E.A., Khan, J., Khan, M.S., Khang, Y.H., Khatab, K., Khater, A., Khater, M.M., Khatony, A., Khazaei, M., Khazaei, S., Khazaei-Pool, M., Khubchandani, J., Kianipour, N., Kim, Y.J., Kimokoti, R.W., Kinyoki, D.K., Kisa, A., Kisa, S., Kolola, T., Kosen, S., Koul, P.A., Koyanagi, A., Kraemer, M.U.G., Krishan, K., Krohn, K.J., Kugbey, N., Kumar, G.A., Kumar, M., Kumar, P., Kuupiel, D., Lacey, B., Lad, S.D., Lami, F.H., Larsson, A.O., Lee, P.H., Leili, M., Levine, A.J., Li, S., Lim, L.L., Listl, S., Longbottom, J., Lopez, J.C.F., Lorkowski, S., Magdeldin, S., Abd El Razek, H.M., Abd El Razek, M.M., Majeed, A., Maleki, A., Malekzadeh, R., Malta, D.C., Mamun, A.A., Manafi, N., Manda, A.L., Mansourian, M., Martins-Melo, F.R., Masaka, A., Massenburg, B.B., Maulik, P.K., Mayala, B.K., Mazidi, M., McKee, M., Mehrotra, R., Mehta, K.M., Meles, G.G., Mendoza, W., Menezes, R.G., Meretoja, A., Meretoja, T.J., Mestrovic, T., Miller, T.R., Miller-Petrie, M.K., Mills, E.J., Milne, G.J., Mini, G.K., Mir, S.M., Mirjalali, H., Mirrakhimov, E.M., Mohamadi, E., Mohammad, D.K., Darwesh, A.M., Mezerji, N.M.G., Mohammed, A.S., Mohammed, S., Mokdad, A.H., Molokhia, M., Monasta, L., Moodley, Y., Moosazadeh, M., Moradi, G., Moradi, M., Moradi, Y., Moradi-Lakeh, M., Moradinazar, M., Moraga, P., Morawska, L., Mosapour, A., Mousavi, S.M., Mueller, U.O., Muluneh, A.G., Mustafa, G., Nabavizadeh, B., Naderi, M., Nagarajan, A.J., Nahvijou, A., Najafi, F., Nangia, V., Ndwandwe, D.E., Neamati, N., Negoi, I., Negoi, R.I., Ngunjiri, J.W., Thi Nguyen, H.L., Nguyen, L.H., Nguyen, S.H., Nielsen, K.R., Ningrum, D.N.A., Nirayo, Y.L., Nixon, M.R., Nnaji, C.A., Nojomi, M., Noroozi, M., Nosratnejad, S., Noubiap, J.J., Motlagh, S.N., Ofori-Asenso, R., Ogbo, F.A., Oladimeji, K.E., Olagunju, A.T., Olfatifar, M., Olum, S., Olusanya, B.O., Oluwasanu, M.M., Onwujekwe, O.E., Oren, E., Ortega-Altamirano, D.D.V., Ortiz, A., Osarenotor, O., Osei, F.B., Osgood-Zimmerman, A.E., Otstavnov, S.S., Owolabi, M.O., P, A.M., Pagheh, A.S., Pakhale, S., Panda-Jonas, S., Pandey, A., Park, E.K., Parsian, H., Pashaei, T., Patel, S.K., Pepito, V.C.F., Pereira, A., Perkins, S., Pickering, B.V., Pilgrim, T., Pirestani, M., Piroozi, B., Pirsaheb, M., Plana-Ripoll, O., Pourjafar, H., Puri, P., Qorbani, M., Quintana, H., Rabiee, M., Rabiee, N., Radfar, A., Rafiei, A., Rahim, F., Rahimi, Z., Rahimi-Movaghar, V., Rahimzadeh, S., Rajati, F., Raju, S.B., Ramezankhani, A., Ranabhat, C.L., Rasella, D., Rashedi, V., Rawal, L., Reiner, R.C., Jr., Renzaho, A.M.N., Rezaei, S., Rezapour, A., Riahi, S.M., Ribeiro, A.I., Roever, L., Roro, E.M., Roser, M., Roshandel, G., Roshani, D., Rostami, A., Rubagotti, E., Rubino, S., Sabour, S., Sadat, N., Sadeghi, E., Saeedi, R., Safari, Y., Safari-Faramani, R., Safdarian, M., Sahebkar, A., Salahshoor, M.R., Salam, N., Salamati, P., Salehi, F., Zahabi, S.S., Salimi, Y., Salimzadeh, H., Salomon, J.A., Sambala, E.Z., Samy, A.M., Santric Milicevic, M.M., Jose, B.P.S., Saraswathy, S.Y.I., Sarmiento-Suarez, R., Sartorius, B., Sathian, B., Saxena, S., Sbarra, A.N., Schaeffer, L.E., Schwebel, D.C., Sepanlou, S.G., Seyedmousavi, S., Shaahmadi, F., Shaikh, M.A., Shams-Beyranvand, M., Shamshirian, A., Shamsizadeh, M., Sharafi, K., Sharif, M., Sharif-Alhoseini, M., Sharifi, H., Sharma, J., Sharma, R., Sheikh, A., Shields, C., Shigematsu, M., Shiri, R., Shiue, I., Shuval, K., Siddiqi, T.J., Silva, J.P., Singh, J.A., Sinha, D.N., Sisay, M.M., Sisay, S., Sliwa, K., Smith, D.L., Somayaji, R., Soofi, M., Soriano, J.B., Sreeramareddy, C.T., Sudaryanto, A., Sufiyan, M.B., Sykes, B.L., Sylaja, P.N., Tabares-Seisdedos, R., Tabb, K.M., Tabuchi, T., Taveira, N., Temsah, M.H., Terkawi, A.S., Tessema, Z.T., Thankappan, K.R., Thirunavukkarasu, S., To, Q.G., Tovani-Palone, M.R., Tran, B.X., Tran, K.B., Ullah, I., Usman, M.S., Uthman, O.A., Vahedian-Azimi, A., Valdez, P.R., van Boven, J.F.M., Vasankari, T.J., Vasseghian, Y., Veisani, Y., Venketasubramanian, N., Violante, F.S., Vladimirov, S.K., Vlassov, V., Vos, T., Vu, G.T., Vujcic, I.S., Waheed, Y., Wakefield, J., Wang, H., Wang, Y., Wang, Y.P., Ward, J.L., Weintraub, R.G., Weldegwergs, K.G., Weldesamuel, G.T., Westerman, R., Wiysonge, C.S., Wondafrash, D.Z., Woyczynski, L., Wu, A.M., Xu, G., Yadegar, A., Yamada, T., Yazdi-Feyzabadi, V., Yilgwan, C.S., Yip, P., Yonemoto, N., Lebni, J.Y., Younis, M.Z., Yousefifard, M., Yousof, H.S.A., Yu, C., Yusefzadeh, H., Zabeh, E., Moghadam, T.Z., Bin Zaman, S., Zamani, M., Zandian, H., Zangeneh, A., Zerfu, T.A., Zhang, Y., Ziapour, A., Zodpey, S., Murray, C.J.L., Hay, S.I., 2019. Mapping 123 million neonatal, infant and child deaths between 2000 and 2017. Nature 574, 353-358.

Fortin, M., Stewart, M., Poitras, M.E., Almirall, J., Maddocks, H., 2012. A systematic review of prevalence studies on multimorbidity: toward a more uniform methodology. Ann Fam Med 10, 142-151.

Galetsi, P., Katsaliaki, K., 2020. Big data analytics in health: an overview and bibliometric study of research activity. Health Info Libr J 37, 5-25.

Hall, M., Laut, K., Dondo, T.B., Alabas, O.A., Brogan, R.A., Gutacker, N., Cookson, R., Norman, P., Timmis, A., de Belder, M., Ludman, P.F., Gale, C.P., National Institute for Cardiovascular Outcomes, R., 2016. Patient and hospital determinants of primary percutaneous coronary intervention in England, 2003-2013. Heart 102, 313-319.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M.L., Goldstein, H., 2017. Challenges in administrative data linkage for research. Big Data Soc 4, 2053951717745678.

Hilbert, M., Lopez, P., 2011. The world's technological capacity to store, communicate, and compute information. Science 332, 60-65.

Jaeschke, R., Guyatt, G.H., 1989. Why different trials on digitalis give conflicting data. Cardiovasc Drugs Ther 2, 727-731.

Johansen, A., Boulton, C., Hertz, K., Ellis, M., Burgon, V., Rai, S., Wakeman, R., 2017a. The National Hip Fracture Database (NHFD) - Using a national clinical audit to raise standards of nursing care. Int J Orthop Trauma Nurs 26, 3-6.

Johansen, A., Tsang, C., Boulton, C., Wakeman, R., Moppett, I., 2017b. Understanding mortality rates after hip fracture repair using ASA physical status in the National Hip Fracture Database. Anaesthesia 72, 961-966.

Jones, K.H., Ford, D.V., Jones, C., Dsilva, R., Thompson, S., Brooks, C.J., Heaven, M.L., Thayer, D.S., McNerney, C.L., Lyons, R.A., 2014. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: a privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform 50, 196-204.

Marengoni, A., Angleman, S., Melis, R., Mangialasche, F., Karp, A., Garmen, A., Meinow, B., Fratiglioni, L., 2011. Aging with multimorbidity: a systematic review of the literature. Ageing Res Rev 10, 430-439.

Metcalfe, D., Judge, A., Perry, D.C., Gabbe, B., Zogg, C.K., Costa, M.L., 2019. Total hip arthroplasty versus hemiarthroplasty for independently mobile older adults with intracapsular hip fractures. BMC Musculoskelet Disord 20, 226.

Mokraoui, N.M., Haggerty, J., Almirall, J., Fortin, M., 2016. Prevalence of self-reported multimorbidity in the general population and in primary care practices: a cross-sectional study. BMC Res Notes 9, 314.

Morris, S., Ramsay, A.I.G., Boaden, R.J., Hunter, R.M., McKevitt, C., Paley, L., Perry, C., Rudd, A.G., Turner, S.J., Tyrrell, P.J., Wolfe, C.D.A., Fulop, N.J., 2019. Impact and sustainability of centralising acute stroke services in English metropolitan areas: retrospective analysis of hospital episode statistics and stroke national audit data. BMJ 364, l1.

Neuburger, J., Currie, C., Wakeman, R., Johansen, A., Tsang, C., Plant, F., Wilson, H., Cromwell, D.A., van der Meulen, J., De Stavola, B., 2017. Increased orthogeriatrician involvement in hip fracture care and its impact on mortality in England. Age Ageing 46, 187-192.

Pawar, S., Liew, T.O., Stanam, A., Lahiri, C., 2020. Common cancer biomarkers of breast and ovarian types identified through artificial intelligence. Chem Biol Drug Des.

Peek, N., Rodrigues, P.P., 2018. Three controversies in health data science. Int J Data Sci Anal 6, 261-269.

Saklayen, M.G., 2018. The Global Epidemic of the Metabolic Syndrome. Curr Hypertens Rep 20, 12.

Smith, M.C., Ben-Shlomo, Y., Dieppe, P., Beswick, A.D., Adebajo, A.O., Wilkinson, J.M., Blom, A.W., National Joint Registry for England, W., Northern, I., 2017. Rates of hip and knee joint replacement amongst different ethnic groups in England: an analysis of National Joint Registry data. Osteoarthritis Cartilage 25, 448-454.

Tucker, T.C., Durbin, E.B., McDowell, J.K., Huang, B., 2019. Unlocking the potential of population-based cancer registries. Cancer 125, 3729-3737.

Vetrano, D.L., Palmer, K., Marengoni, A., Marzetti, E., Lattanzio, F., Roller-Wirnsberger, R., Lopez Samaniego, L., Rodriguez-Manas, L., Bernabei, R., Onder, G., Joint Action, A.W.P.G., 2019. Frailty and Multimorbidity: A Systematic Review and Meta-analysis. J Gerontol A Biol Sci Med Sci 74, 659-666.

Wallace, S.E., Gaye, A., Shoush, O., Burton, P.R., 2014. Protecting personal data in epidemiological research: DataSHIELD and UK law. Public Health Genomics 17, 149-157.

Wolff, J.L., Starfield, B., Anderson, G., 2002. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. Arch Intern Med 162, 2269-2276.