Code of Practice for the conduct of research using data held in CHC Trustworthy Research Environments

Version 1.0, 10th June 2019

Editor: Prof Dipak Kalra, Director of External Affairs, on behalf of the CHC

Contents

Code of Practice for the conduct of research using data held in CHC Trustworthy Research Environments

Purpose

This proposed Code of Practice defines the acceptability criteria and governance rules for the conduct of research using data held within Connected Health Cities Trustworthy Research Environments, which might be undertaken by public or private research organisations. It is being more widely shared to encourage review and feedback from other health data research communities, with the eventual aim of encouraging UK and European convergence on a common code of practice and rules for reusing health data for research.

Scope within Connected Health Cities

The four city regions who form the Connected Health Cities have each established mechanisms and permissions to generate repositories of anonymised data derived from health and care organisations, patients and citizens in their part of the north of England. In some scenarios a generic de-identified repository, known as an Ark, is being created that acts as a master data repository from which specific research-relevant data sets can be derived and accessed for approved research purposes. In other scenarios research-relevant data sets will be created as needed by direct extraction and anonymisation from one or more health and care organisational record systems. A further scenario is that a regional Health Information Exchange, created to serve continuity of care across health and care providers, acts as the data source for generating anonymous research data sets. For all of these scenarios, the end result is the creation of a repository that contains the data set needed to conduct a specified research study, and which is placed in a secure environment for access by approved research users. This repository infrastructure (with governed access) is known as a Trustworthy Research Environment (TRE). It is the intention that TRE data should be used in specified ways for a set of specific purposes ranging from quality improvement across the north of England and for research (i.e. fostering Learning Health Systems), all with the end goals of improving the health of the population and contributing to wider knowledge. These organisational scenarios are described in Appendix 1.

For each approved research study, the relevant TRE custodians will (separately or jointly) establish one or more *TRE instances* (e.g. virtual machines), containing an agreed extract of their health and care data, anonymised to declared standards (see Appendix 2), as the data set required for that particular approved research study. Researchers will only have access to the anonymised data contained in their nominated TRE instance.

There is a need to establish a formal basis on which research uses of TREs by external organisations are permitted, with corresponding terms and conditions. The agreed final version of this Code of Practice is intended to define this basis and the (governance) terms, with the aim of establishing a common position that is endorsed by all four CHC regions and operated via a joint *Research Access Governing Board (RAGB)*, which includes membership of all four regions so that final decision making on the approval of every research study

respects the autonomy of each region. The constitution of this RAGB will include other stakeholders, including patient and public representatives, described in Appendix 1. The rest of this document defines these terms and governance terms (rules).

The members of the CHC who have contributed to this report and its preparatory meetings are listed in Appendix 3.

Acceptable research use of TRE data

It is of paramount importance to assure TRE custodians, their data contributing stakeholders and the public, that TRE data is only used for acceptable purposes. This Code of Practice prioritises the purpose of the intended research study as a critical acceptance criterion for approving the use of TRE data. The Code of Practice proposes the concept of *bona fide research*, which has precedent in prior work endorsed by the MRC and by European R&D projects.

The over-riding objective of *bona fide research* must be to discover new knowledge and learning aimed to improve the health of society, intended to enhance the wellbeing and health of all citizens and those involved in the delivery of healthcare, and intended for the public good and to be made publicly accessible (i.e. published) without undue delay. (The NHS more specifically states that NHS data should be targeted at improving health outcomes, which is consistent with the previous sentence.)

It is recognised that there are interim/provisional stages of the research process which might not lend themselves to public consumption. Examples include theory generation, procedural development and feasibility. It is also accepted that at times the purpose of research is to corroborate existing findings, or to re-examine prior findings. Although the analysis of findings is sometimes not publishable/usable, the research organisation must still be accountable to the RAGB for research that has been undertaken using TRE data and produce a publicly-accessible summary of the investigation that was undertaken.

Other principles which the consultative group felt important to emphasise, even though their assessment is more subjective, are that the research should align with human values and be compatible with human ideals of human dignity, rights, freedoms and cultural diversity. The research should not exhibit discrimination, but favour and help achieve health and care equity. The research should respect, support and aim to improve health and social care services. The research should aim for benefit to a broad population (i.e. applicable to as many people as possible) rather than to an individual or an organisation. Economic prosperity should not be the main aim of the research: financial gain should aim to function as a means of continuing beneficial research and the delivery of safe and effective healthcare solutions into health and social care. Principles should also consider that increasing business opportunities benefits regional communities and potentially accelerates innovation as well as enhancing skills and knowledge.

The Data Sharing Agreement that covers the data coming into the TRE must specify the defined bona fide use of the data.

HRA approval is not required for the establishment of research databases but TRE data custodians may apply on a voluntary basis for ethical review of the arrangements for collection, storage, use and distribution of data, including arrangements for release of non-identifiable data for analysis by researchers. For access and processing the identifiable data without consent Research Ethics Committee (REC) approval is required by law, and an application to the Confidentiality Advisory Group under Section 251 of the NHS Act 2006 to set aside the common law duty of confidentiality owed by care professionals to their patients or clients.

Once complete, research should be published in a timely manner. Results should not be withheld to maximise financial or reputational gain. Research results should be published or made accessible for the benefit of all whether deemed by the researcher as successful or unsuccessful. Research findings should be shared with the public in clear jargon-free language.

A culture of trust, transparency and mutual support is vital between researchers, data controllers and data subjects, who should co-operate to prevent, investigate and mitigate potential malicious use of data.

Examples of bona fide research purposes

The consultative group considered carefully whether a definitive list of categories of acceptable research could be defined but concluded that the innovative nature of research in health meant that a closed list of acceptable categories would prove unworkable. However, the following examples were collated to serve as a guide to the RAGB and might be presented to the public as examples of the kinds of research for which north of England health data could be used – although these are not exclusive.

- Develop new treatments: e.g. clinical trial feasibility, patient recruitment
- Derive evidence of outcomes and effectiveness
- Monitor and improve health outcomes
- Risk and health needs assessment
- Derive evidence for regulatory and HTA (e.g. NICE) submissions
- Enhance understanding of disease, progression, current standards of care
- Conduct pharmacovigilance studies, monitor patient safety
- Profile biomarkers and target populations for therapies (precision medicine)
- Epidemiological measurement e.g. estimating disease prevalence or survival rate
- Model and evaluate patient pathways
- Explore potential for improvement in patient health
- Technology development within the healthcare arena e.g. medical devices, sensors, wearables
- Development and validation of algorithms and Artificial Intelligence
- Preliminary product design

The following examples were collated of the kinds of research which the consultative group felt would be **undesirable** to support and would present a reputational risk to north of England health and care organisations if conducted.

- Research that has been refused ethical approval, if it was required for a particular study (although this is not required for research only using anonymous data)
- Weapons development and research, including development of biological weapons (although research into treatments following biological attack may be acceptable)
- Drugs for use in capital punishment, interrogation or torture
- Eugenics
- Political analyses where there is party political gain motivating the research
- Discrimination (although it may be acceptable to conduct population profiling to assess the equity of a care service, to biologically target appropriate therapies and to assess health risks)
- Marketing of an existing product (although it would be appropriate to conduct usability testing of devices or uncover unmet treatment needs)
- Research where the sole outcome is a financial interest
- Research which would be illegal in this and/or perhaps in other countries

This list is also intended as an informal guide to the RAGB. However, the consultative group suggests that this list might not be appropriate to publicise. Assessing bodies such as the RAGB should have enough information to enable them to perform a risk assessment that proposals are ethically sound and in keeping with the principles of improving health and well-being.

Acceptable research organisations

It is widely recognised that distinguishing the capability and acceptability to undertake research on the basis of whether a research organisation is publicly or privately funded fails to recognise that company-sponsored and/or company-undertaken research frequently leads to innovative products, such as new or improved medicines and devices, that benefit the future provision of healthcare and therefore ultimately benefit the public. The consultative group examined this matter in detail and concluded that it is more important to base a research access decision on the suitability of an organisation to conduct *bona fide* research rather than how the organisation is financed.

This Code of Practice therefore advocates making research access decisions on the basis of the following definition of a *bona fide research organisation*.

A bona fide research organisation is one that is appointed or accredited or funded to undertake bona fide research and has made public its commitment to adhere to recognised research governance principles (such as Good Clinical Practice[1]). It is not a requirement that such research is the primary business of that organisation, or that all of the research undertaken by that organisation is published. It is not a requirement that the organisation be publicly funded.

Examples of organisations that would usually be permitted research access to north of England health data:

- Health and social care provider
- Academic research organisation (e.g. university)
- Public health organisation
- Healthcare funder (health ministry, commissioning group, health insurer)
- Patient association or charity
- Regulatory body (e.g. MHRA, NICE)
- Pharma company, biotech company, AI company
- Manufacturers of medical devices, appliances, systems etc.
- ICT (software, platform) developer or service provider, digital therapeutics

The consultative group recommends that neither a list of acceptable organisations nor of unacceptable or blacklisted organisations be published, but that research requests should be assessed on a case-by-case basis. The research track record of any requesting organisation, such as any recent history of data protection or research conduct breaches, will be taken into account. (A cumulative list of organisations to whom data access has historically been granted will be published as part of transparency.)

Bona fide research user

A bona fide research user is a person working within or for a bona fide research organisation whose contract of employment or service contract or student status permits such research

[1] Published by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). https://www.ich.org/home.html

and requires appropriate governance in the conduct of that research, including governance of their handling of information. Research users must be working within or contracted to an approved (bona fide) research organisation, and not be an independent person. Suitable training in data protection and information security must have been given, and regularly updated. He or she must have an employment contract (or be otherwise bound by) a policy which identifies the responsibility to protect data subject privacy, linked to disciplinary action should they fail to do so.

Researcher users may only execute analysis queries for an investigation that conforms to an approved purpose.

Researcher users may only use data set extracts (if they are permitted to directly access anonymous subject level data from the TRE) within their own research organisation, or joint research project, with staff and contractors involved in an investigation conforming to an approved purpose. Normally disclosure controls will restrict the results that could be taken out of the TRE.

The legal basis for the use of north of England health data for research

A legal basis is required, under UK Data protection legislation and the EU General Data Protection Regulation, for all processing of personal data[2]. (Processing in this context includes the collection, storage, sharing, analysis and other transformations of personal data.) There is still national and European debate about the circumstances in which a legitimate interest legal basis, a public interest legal basis, or informed consent, should be used as the legal basis for research conducted on "big" data repositories. Scientific purposes are recognised as a legitimate basis for processing personal data, but the details of how this purpose is interpreted are also still not clear. (Note: Public interest is the general welfare and rights of the public that are to be recognised, protected and advanced. Disclosures in the public interest based on the common law of confidentiality are made where disclosure is essential to prevent a serious and imminent threat to public health, national security, the life of the individual or a third party or to prevent or detect serious crime. This basis does seem applicable to the conduct of research.) However, it is important that guidance on this issue is updated in a timely way to reflect legal changes.

Anonymised data

Data protection legislation, and therefore the requirement for a legal basis, is not applicable to data that have been anonymised. Being anonymised means that the party in possession of data, or accessing the data (e.g. remotely), has no means of reidentifying any individual in the data set, even if combining the data with other data that is readily available to them.

There is no requirement in UK or European data protection legislation for data subject consent for the processing of data that have been anonymised. (Consent or another legal

[2] There is a public interest test applied to all research projects by the UK Statistics Authority. (Their code of practice is not for specifically for health data.) https://www.statisticsauthority.gov.uk/wp-content/uploads/2018/08/COP_Research-and-Accreditation_A4.pdf

basis will be needed to perform the anonymisation, since personal data must be processed to create an anonymised extract, but this basis is outside the scope of this Code of Practice.)

Anonymisation is a relative rather than absolute process: a small possibility might remain in an anonymous data set that some individuals can become recognisable through some of the data items in the set, possibly through pattern matching with some other accessible data. There are therefore guidelines on what constitutes adequate anonymisation, which are summarised in Appendix 2.

This Code of Practice recommends that, at present, data transferred into Trustworthy Research Environments (TREs) for research use are de-identified and that data sets made available for research are anonymised to recognised standards, which the RAGB should specify on the basis of nationally or internally endorsed methods (see Appendix 2). This anonymisation must be undertaken by data controllers with permission to process the personal health data before transferring the data directly into a TRE instance, or into an Ark that will later become the source of data for TREs.

Pseudonymised data

Anonymisation may at times make the data unsuitable for a specific research study. Anonymisation makes it impossible to link data in one data set to another (or across a dataset that has been split into multiple tables), since the individuals in each data set cannot be matched to their counterparts. It is also impossible to contact those data subjects to request additional data or bio-samples from them. It may therefore at times be necessary to make a pseudonymous data set available for research.

The EU GDPR considers pseudonymised data to be personal data. The legal basis for conducting scientific research on pseudonymised data is still being debated at EU and country specific levels. Please see a discussion of pseudonymisation, and references to further reading, in Appendix 2.

This Code of Practice recommends that, at present, research access is not granted via TREs to pseudonymised data (unless particular ethical approval is obtained to access it e.g. consent for consent studies that retrieve individualized records to understand an epidemiological issue). This recommendation should be revisited once UK recommendations have been clarified on the legal basis for this kind of data, on the recommended practice in performing pseudonymisation and on the safeguards that are sufficient.

There are also challenges with undertaking research into rare diseases, where it may be difficult or impossible to apply robust anonymisation techniques because individuals are so readily recognisable or discoverable from their “anonymous” clinical picture. At this stage it is recommended for rare disease research that cannot use anonymous data to only be undertaken with ethical approval.

Should research findings indicate a new risk to specific subgroups of patients, then this must be communicated to the relevant care provider organisation(s) who will be responsible for analysing their identifiable data repository to identify and contact those patients, if appropriate. Those data controllers with permission to process the identifiable original data that was exported into a TRE will have the means to re-query their own source data to

identify relevant individuals, such as individuals with a newly-discovered risk. They therefore can, if necessary, make contact with the pool of patients who are presumed to be data subjects within the anonymous research data set. They could also undertake data set linkage on behalf of a research user, before regenerating an enriched anonymous data set for a TRE.

Applicable CHC Governance Principles

The following principles, which are an adapted extract of CHC governing principles for Arks themselves, are applicable to this Code of Practice.

- Ark and TRE controllers must apply technical and physical controls to the processing of data by research users to minimise the risk of unlawful access, data loss and hacking.
- Query access only via TREs will be provided to researchers, not data sets.
- Data access rules and processes must be published, implemented and monitored.
- All requests for data must be assessed by the Research Access Governing Board.
- Organisations requesting data access must be vetted to check that they are capable of protecting their access appropriately.
- Requests for data access must be risk assessed to ensure they are non-identifying.
- Approved data access decisions must be published[3].
- Contracts with research users must make clear their responsibilities to protect data and act in a responsible way.
- Agreements must make clear that the Research Access Governing Board will require the relevant data controller(s) to report to the ICO anyone that deliberately attempt to re-identify individuals.

[3] Please see Appendix 1 describing the proposed transparency of RAGB decisions.

Code of practice for health and care providers and Health Information Exchange controllers

Health and care organisations and/or Health Information Exchange (HIE) controllers may only contribute data to an Ark or a TRE if this is locally approved, and consistent with their terms of reference, ethical approval if required (and participant consent if applicable) and in line with their privacy notice in order to meet transparency requirements of GDPR.

Health and care and/or HIE data controllers are responsible for specifying any constraints that apply to particular data sets which may limit which organisations may reuse the data and for which purposes[4].

Health and care and/or HIE data controllers shall exclusively determine which parts of their total data holding within an HIE will be imported into an Ark or a TRE and may vary this specification over time (including withdrawal).

Health and care and/or HIE data controllers retain the capability to identify individuals, but Ark/TRE controllers may not have this capability.

Health and care and/or HIE data controllers may have the ability to perform subject linkage across their internal databases and with local third parties. They may have the capability to link successive Ark/TRE updates longitudinally. An Ark/TRE custodian should only undertake this if it is an approved and contracted data processor, or can do so via robust privacy protected linkage.

Health and care and/or HIE data controllers will perform de-identification or anonymisation of all health data being imported into an Ark or a TRE. The data linkage, de-identification and anonymisation methods used will be transparent, documented and will conform to minimum standards agreed across the CHC, and nationally.

The act of anonymisation should be performed in systems and by people who are allowed to hold the identifiable data.

Health and care and/or HIE data controllers and Ark/TRE custodians must agree if pseudonymous key-codes will be provided, to whom and under what rules of use, so that future transfers of data will be linked to existing data, and what mechanisms will be used to enable this without prejudicing the anonymous nature of Ark/TRE data.

Code of practice for Ark and TRE custodians

Ark/TRE custodians must confirm that the anonymisation they undertake complies to a nationally or internationally recognised standard.

Ark/TRE custodians must ensure that each approved research study request is consistent with their terms of reference and has ethical approval if applicable.

Ark/TRE custodians must ensure the anonymisation of subject level data before releasing any data set into a TRE.

[4] As an example, data sets obtained from NHS Digital may have constraints on how that data may be used

In order to ensure valid inferences are made from the data, research users must be provided with information about the anonymisation techniques that have been applied to specific data items within the data set, such as how k-anonymity has been applied.

Anonymised data should be treated as if it still carries a small residual risk of re-identification, and therefore still be subject to robust information security practices.

Ark/TRE custodians must ensure that data access agreements and any accompanying policies signed with research organisations obligate the organisations to apply acceptable security and confidentiality measures to the TRE, at least as stringent as those applied to the corresponding data by the Ark/TRE custodians.

Data access agreements must specify:

- the data items and data sources that will be included in a TRE
- any documentation that explains the data that might also be provided, including details of the anonymisation process used
- the time period for which TRE access will be granted
- if data in the TRE will be periodically updated, with what information and how frequently

Data access agreements must specify the access fees that are payable, to whom and when with transparency on what these charges cover. The principles for cost recovery should be consistent across the TREs, including offsetting any third party costs such as those levied by NHS Digital, even if different charging models are applied for data access.

TRE instances must be created to provide access to, in an anonymised form, the necessary data for an approved research study, and external access to the TRE instance be restricted to personnel nominated by the relevant research organisation as having a contractual basis for undertaking the approved research using that data set.

TREs must provide suitable physical and technical information security including audit of the data access and processing functions performed.

Remote access to TRE data must only be provided when appropriate protection can be assured of the remote access channel and of the information security provisions of the accessing organisation.

Autonomy of Ark/TRE custodians

Although this Code of Practice aims to define the rules for research access to all CHC Arks and TREs, it is recognised that each city region may have additional constraints and acceptance criteria that may at times require them to opt out of providing data access to a specific research study. This Code of Practice therefore stipulates that representatives of the Health and care and/or HIE data controllers:

- may set restrictions that are additional to this Code of Practice on which kinds of bona fide user, bona fide purposes their data may be used for (which the RAGB will need to apply)
- may restrict use of different parts of their data holdings for different research purposes, usually by specifying what data may be added to the TRE created for that study
- will always have access to the approved set of research users and research protocols

- may veto any specific research protocol or research user[5]
- may at any point demand that a particular research use be investigated, if deemed necessary, by external auditors
- will be regularly consulted on the operation and evolution of this Code of Practice and of the RAGB

[5] It is proposed that this veto is normally exercised by the city regional member on the RAGB, with the option to consult other stakeholders if needed.

Code of practice for research users

The following sub-sections should be reproduced or referenced within each Data Sharing Agreement that is established between a TRE custodian and a research user organisation.

Legal compliance

Research users must comply with all relevant national and EU data protection and clinical research legislation, and relevant European guidelines, as applicable to their role and purpose of use.

The research user organisation must ensure that any relevant ethics committee approvals have been obtained for the intended purposes of use and locations of accessing the data set (if remote access is permitted) and take responsibility for obtaining any additional approvals that are found necessary.

Purpose limitation

The research user organisation must only use data provided through the TRE for the agreed purposes.

The research user organisation must ensure that all staff involved in processing the data are aware of the purposes for which the data may be processed and of any other constraints and stipulations within the data access agreement that may impact on the processing actions they undertake.

Analyses on the data set must only be performed in order to further those approved purposes. The research organisation must seek permission from the RAGB for any additional processing purposes of the same data set.

A research organisation may be required to demonstrate, through an arranged inspection of audit logs and other relevant documents such as output reports, that data have only been used for the purposes for which access has been granted.

Data protection

The research user organisation must ensure that appropriate information security and physical protection measures are contracted and have been implemented to safeguard the dataset and to prevent unintended disclosures of the data or damage to its integrity.

The research user organisation must ensure that all staff involved in processing the data are adequately trained in information security and privacy protection.

Research users must not seek to re-identify any individuals within an anonymous dataset.

Research users are required to report to the RAGB and relevant TRE custodians any material issues that they discover with the data set that poses a risk to privacy protection or to the health of the data subjects.

The research user organisation must assume that the data, and all rights to use the data, are non-transferable unless terms for onward data sharing with other parties have been pre-agreed with the TRE custodians and are formally specified in the data access agreement. The research user cannot disclose the data set beyond pre-agreed recipients.

Researchers and Ark custodians must be clear about what can and cannot be taken out of the Ark environment.

Publication of results

The research user organisation must attempt the timely publication of research results through a public access channel (though not necessarily free to access), either via a recognised publisher or via an organisational web site, or other agreed output. If publication is not achieved within a certain time-frame – measures must be taken to provide a summary of the findings in a public domain.

Co-authorship arrangements of Ark/TRE, health and care, or other staff, and/or acknowledgements to be included in a publication, must be specified in the data access agreement or by subsequent written mutual agreement.

For research intended to be incorporated into products or services, rather than openly published, that intention must be explicitly stated and formally agreed before data access is granted.

Sanctions

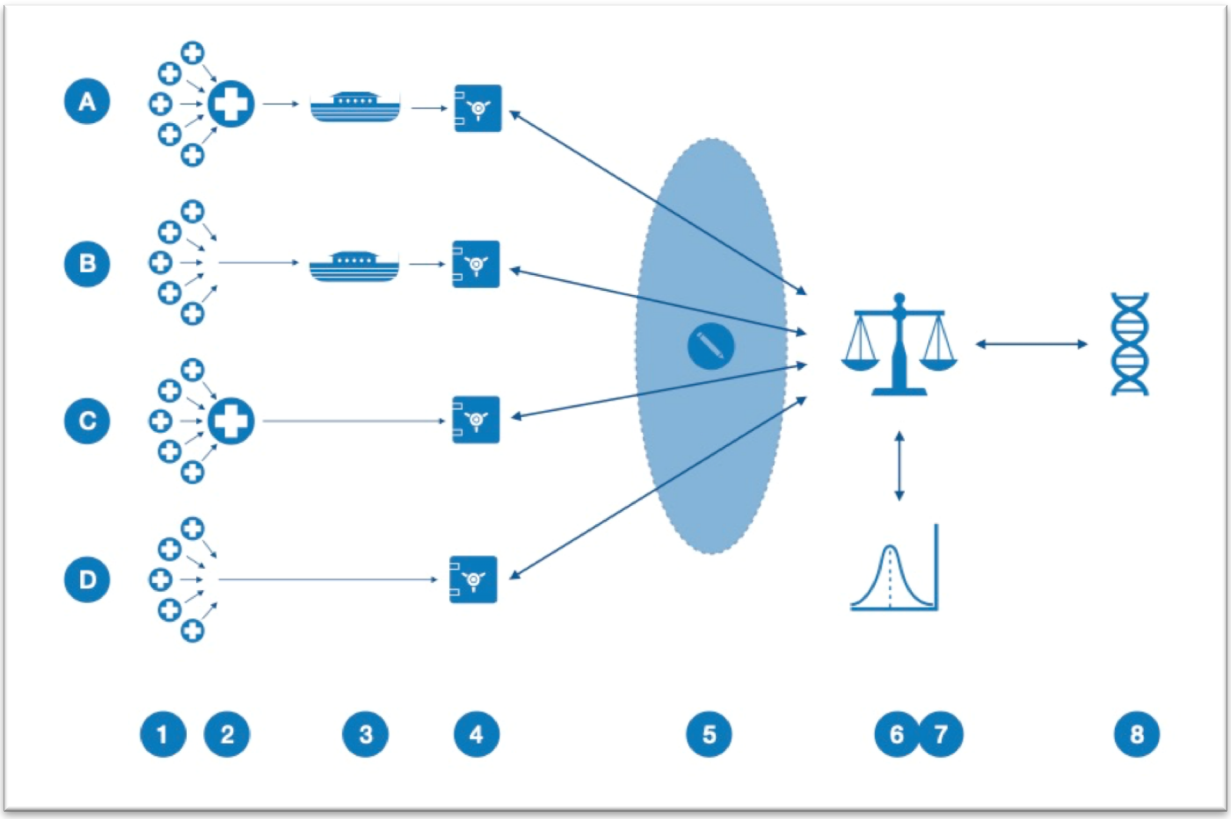
The research user organisation must ensure that appropriate disciplinary measures and sanctions can be applied in the event of staff, students or contractors behaving inappropriately with the data. This may include advising personnel in advance that such measures may be applied and including cautionary terms within employment contracts. This should also include advising personnel that audit logs of activity are kept and regularly inspected.

Research user organisations found to be in breach of this Code of Practice, to have breached the terms of data access agreements, to have used data sets beyond the scope of agreed purposes or to have taken insufficient measures to protect the data and the privacy of individuals within it, may be reported to regulatory authorities such as the UK Information Commissioner's Office.

Appendix 1: Operating model for the Research Access Governing Board

As stated earlier, the four CHC city regions have each established mechanisms and permissions to generate repositories of de-identified data derived from health and care organisations, patients and citizens in their part of the north of England. In some scenarios a generic de-identified repository, known as an Ark, is being created that acts as a master data repository from which specific research-relevant data sets can be derived and accessed for approved research purposes. This Ark might be populated from a regional Health Information Exchange repository or directly from multiple healthcare provider systems. In other scenarios research-relevant data sets will be created as needed by direct extraction and anonymisation from a Health Information Exchange or from one or more health and care organisational record systems. For all of these scenarios, the end result is the creation of a repository that contains the data set needed to conduct a specified research study, and which is placed in a secure environment for access by approved research users.

The figure below illustrates the operating model for research access to CHC health data, according to these four scenarios (A, B, C, D). The small blue crosses are individual healthcare provider organisations, the large blue cross is a Health Information Exchange. The ark symbol is hopefully clear. The safe symbol has been used to depict a TRE. The numbered steps are explained further below.



Four example inter-organisational scenarios are shown in this figure: A-D.

A. In scenario A the health care, social care and voluntary organisations in that region have co-developed or procured a Health Information Exchange (HIE) to allow them to share personal health data to support better continuity of care. They have also agreed to contribute much of this data, in anonymous or pseudonymous form, from the HIE to a regional Ark which will hold the data in anticipation of future research uses. For each approved study the Ark custodian will create a TRE with the data relevant to that study.
B. In scenario B, the health care, social care and voluntary organisations in that region contribute anonymous data directly to their regional Ark but have not elected to establish an HIE.
C. In scenario C the health care, social care and voluntary organisations have created an HIE, and the HIE controller directly populates TREs with relevant anonymised data as needed. There is no generalised Ark-like research data repository.
D. In scenario D there is neither an HIE or an Ark, and so a set of health care, social care and voluntary organisations in that region will directly populate a TRE with relevant data on an as needed basis.

The numbered points immediately below the figure are described here.

1. Several different health care, social care, voluntary organisations and patients agree to contribute health data for reuse, for Learning Health System purposes. They might optionally contribute this data via a Health Information Exchange (HIE) operated by their city region, which is established primarily to support continuity of care and local quality improvement. HIEs will contain personal health data that is fully identified and used for care delivery. Any arrangements made for anonymised or pseudonymised use of the data for local quality improvement purposes are outside the scope of this Code of Practice.
2. The care organisations or HIEs are responsible for formalising agreements to contribute part or all of this data to their local Ark, if this is established. The care organisations or HIEs are responsible for undertaking any longitudinal linkage or linkage across care organisations, and responsible for undertaking anonymisation or pseudonymisation of the data before it is transferred to the Ark. Ark custodians may perform this if permitted as data processors or via privacy protected linkage. The care organisations, the HIEs and the Arks are jointly responsible for ensuring that the necessary data sharing and data protection assurances and agreements are in place for this.
3. Ark custodians are responsible, on behalf of the contributing health and care organisations, for agreeing acceptable research purposes and research organisations which may use the data, in an anonymous form, only for uses approved by the contributing health and care organisations. Ark custodians are responsible for obtaining any necessary approvals for the research which they agree to support. Ark custodians are responsible for complying with data protection legislation if they hold personal data (such as pseudonymous data). Ark custodians are responsible for creating anonymous TRE instances that contain only the data necessary for conducting each approved research study. Ark custodians must communicate any

specific constraints that supplement this Code of Practice to the Research Access Governing Board.

4. If an Ark is not established in a city region, the HIE or the care organisations creating a TRE directly will need to take the responsibilities indicated for Arks in the point above.
5. If agreed, a single nominated legal entity may act as the joint contracting body (prime contractor) for research using TRE data on behalf of the four regions, with which it may hold a framework agreement with each region or may create an agreement for the specific data access arrangements for each study, as appropriate. If it is decided not to nominate a single legal entity to act on behalf of the regions, then contracting responsibilities will be undertaken by each region involved in a research study.
6. The Research Access Governing Board is responsible for acting as a common broker to present the opportunity for research access to external bona fide research organisations, whilst respecting any specific constraints on research access that have been specified by individual regions. It is responsible for verifying bona fide research organisation status and verifying the suitability of the purpose of any intended research. Through its regionally-appointed members, it will verify the approval from of each region that has relevant data to the research or note an opt out by any region to a particular study. This Board is responsible for providing single access channels for research organisations to find out about the available data, and to submit data access requests and research protocols. It is also responsible for summarising information about the research use made of data and its own governance activities, for the public. It will investigate any data-related concerns and issues if they arise.
7. Where necessary, the governing board will refer submitted protocols to a scientific advisory group for appraisal. This group will need to be established.
8. Bona fide research organisations will have access to a single channel and process for submitting data access requests, for making preliminary enquiries, for agreeing to the stipulated rules for agreed data access and for demonstrating compliance to this Code of Practice.

The Research Access Governing Board will therefore need to:

Manage research access:

- maintain a publicly accessible catalogue of the kinds and scale of north of England health data available for research
- support a single point of enquiry and access to submitting research access requests
- verify the bona fide nature of proposed research studies and research organisations
- assess and approve in principle requests for specific research protocols to be conducted on north of England health data
- refer research protocols to a scientific advisory board, if appropriate, to verify the scientific, informatics and statistical aspects of the proposed study
- require that research protocols have been referred to and approved by a Research Ethics Committee, if appropriate

- refer endorsed studies to the joint contracting body, if this is established, to formalise the data access contract and terms; if not, to refer the endorsed studies to each region to formalise through a data sharing agreement
- verify appropriate research conduct including the outcome and publication of research results
- verify that agreed remunerations or fees have been paid to TRE controllers for the use of data, if this is not undertaken by a separate joint contracting body

Govern research access:

- oversee the overall use of data within and across the TREs by researchers, and highlight any issues or difficulties faced by research organisations and/or TRE controllers or their technical staff
- maintain a publicly accessible inventory of active and completed research studies, including which regions have contributed data to them
- present summaries of TRE use, and the anticipated societal benefits from the research undertaken, to the public and to relevant organisations
- investigate areas of concern about a research study or organisation, reporting to research funding sponsors if necessary
- ensure fair conduct by TRE custodians
- require that relevant parties are notified about any suspected data breaches, and obtain evidence that this has been done
- operate and evolve this Code of Practice, especially in the light of emerging legislation and guidance regarding implementation of the EU GDPR

The Board will need to include representatives of all collaborating regions, representatives of patients and the public, and expertise that includes research ethics, data protection, information security, information architecture and contracting. Membership may include nominated health and care organisations in the north of England, and public and private research user communities, possibly as observers. Provision should be made to co-opt additional areas of expertise on a standing or ad hoc basis.

Note that it is not proposed that this Board provides a review of the scientific relevance and formulation of the proposed research or of the suitability of the available health data to the research question, nor as a Research Ethics Committee. This Board may require evidence that other bodies have considered these issues, where relevant.

Formal terms of reference for this governing body, including a more complete list of roles and responsibilities, will be defined later in 2019.

Code of practice for the Governing Board

The Research Access Governing Board must publish transparent information about the procedure (including criteria and priorities) for evaluating and deciding on research protocols and data access requests.

The criteria for reviewing a data sharing request must take into account:

- whether the purpose of the proposed research is bona fide, is consistent with the TRE custodian's terms of reference, and any applicable ethical approval or participant consent (if this is required)
- whether the requesting organisation is a bona fide research organisation, with an acceptable track record of research conduct
- whether the relevant data are held and available for use, and consistent with any specific constraints or rules of the individual TREs involved
- whether a risk assessment identifies an unacceptable risk to the confidentiality of the participants' identities or if disclosing inferences might be made
- whether a risk assessment identifies an unacceptable risk to the reputation of the TRE controller or the cohort of data subjects.

Timely feedback must be provided to data access requesters explaining the outcome of their request and what courses of action are open to them. For declined requests, the grounds for rejection must be explained and options for re-submission must be specified.

For approved data sharing requests, any funding implications for data use and collaboration must be stated explicitly.

The Board must ensure that any specific constraints or rules set by individual regions are upheld when making recommendations for approval of a submitted research protocol.

The Board must publish information about its activities, and the research conducted on north of England health data, in accordance with this Code of Practice.

Transparency of the Research Access Governing Board

In the interests of ensuring public trust and acceptance in the use of north of England health data for research, it is recommended that the Research Access Governing Board maintain the following information on a public web site.

- The nature and purpose of TREs, data feeds, anonymisation measures normally used, technical security measures and audit trails (using lay terminology)
- The principles for granting data access
- This Code or Practice and a sample Data Processing Agreement template
- The business model framework for charging fees for data access, and how those funds are distributed, ideally with annual accounts
- Information about the Research Access Governing Board: its membership, terms of reference, periodic summaries of its activity, minutes of its meetings (excluding any confidential items discussed)
- An agreed non-confidential summary of approved research protocols (fuller details may be made public after an agreed time limit)
- A public summary of results from approved research
- An anonymised list of declined requests, with outline reasons

It may also be useful to list the public bodies and companies that make use of north of England health data or who have joined a CHC user network if this is established. For each organisation, to provide:

- an outline of their product areas relevant to north of England health data use

- an outline of why they wish to use north of England health data in the near future (high level, respecting commercial confidence, but focusing on the health and care benefits)
- an indication of how they foresee health and care benefits from their use of the data and/or their final products (high level)

It is not recommended to make public the following kinds of materials that the Research Access Governing Board may handle.

- Minutes of the Research Access Governing Board relating to confidential matters
- Company specific policies regarding compliance with the north of England health data access and use policy (such as staff members with access rights)
- Detailed research protocols and data analysis protocols submitted by research organisations
- Correspondence and clarifications from companies about data/results handling for specific investigations
- Correspondence related to obtaining ethical approvals
- The outcome of any investigations into concerns

Business Model

This Code of Practice does not define the business model for operating research access to north of England health data across the four regions. However, it defines the following principles that the business model should uphold.

External organisations should normally be charged a fee for making use of north of England health data. This fee should take into account

- the scale of the proposed research and of the data that is to be accessed
- the number of regions whose data needs to be accessed
- any special work needed by TRE personnel to link, clean, cross map or otherwise process the data to optimise its suitability for the intended research
- the anticipated business value to the research organisation for conducting the research, which may include taking into account whether the organisation is publicly or privately funded
- whether a fee exemption should be granted to some kinds of organisation, such as health care organisations or patient charities

It is anticipated that a high proportion of the fee income generated from data access, especially in the early years, will be used to offset the operating costs of the TREs and of the central governing and contracting bodies responsible for brokering and overseeing research data access.

It is strongly recommended that any net profit from research data access is primarily directed towards health and social care organisations contributing data to the Arks. It is recommended that a distribution formula be agreed at the outset, even if initially hypothetical – to assure public support, and be reviewed annually. This distribution formula may need to take into account the relative volumes of research data usage across the north of England, and the proportion of each region's data contributed by the different health and social care organisations.

Appendix 2 Anonymisation and pseudonymisation of data

This is a high level summary of the basic concepts. More detailed explanations and examples of anonymisation and pseudonymisation are widely available.

Anonymisation

Anonymisation is the process of removing or modifying the values of certain parts of a dataset so that it is not possible (using means reasonably likely to be used) to identify any individual in the dataset, i.e. the data should be permanently non-attributable to an individual. Data which is rendered anonymous is outside the scope of European data protection laws.

Commonly used methods include:

- completely deleting data items (fields) that incorporate demographic descriptors such as personal names and addresses;
- generalising (blurring) values to make them less distinctive but retaining research value in the data, such as converting a date of birth to a year of birth or five-year age band, or converting a precise occupation to an occupational category;
- deliberately modifying values to make them no longer precisely correct but good enough for the intended research, such as modifying a clinical appointment date or an operation date to another date in the same month;
- selecting only a small proportion of records e.g. a sample of 1% of health records from the population;
- replacing a laboratory result with a range, such as a CD4 count being < 250 rather than 241;
- carrying out more sophisticated modifications to multiple values based on the statistical likelihood of a value pattern being distinctive (such as k-anonymity).

Good anonymisation codes are available from the UK Information Commissioner's Office[6] and the UKAN Anonymisation Decision Making Framework[7].

However well anonymised, health data about individuals may carry some residual risk of identifying particular individuals or of the attribution of particular characteristics to recognisable population subgroups. Even aggregate data may carry some risk of identifying individuals, in particular if the data have been aggregated from a narrowly defined population, for example as occurs in the study of rare diseases.

The CHC Information Governance Working Group is working on good practice guidelines on anonymisation, which will be updated once the UK Information Commissioner's Office has updated its national guidelines on this. Certain aspects of good practice, such as small cell size suppression, are still being discussed.

It may also be relevant to note that the Article 29 Working Party highlighted the relative (dynamic) nature of anonymisation with respect to computing power and the wider availability of data via the internet.

[6] https://ico.org.uk/media/1061/anonymisation-code.pdf
[7] https://ukanon.net/ukan-resources/ukan-decision-making-framework/

Article 29 Data Protection Working Party

Opinion 4/2007 on the concept of personal data[8]

Means to identify

Recital 26 of the Directive pays particular attention to the term "identifiable" when it reads that “whereas to determine whether a person is identifiable account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person.” This means that a mere hypothetical possibility to single out the individual is not enough to consider the person as “identifiable”.

If, taking into account “all the means likely reasonably to be used by the controller or any other person”, that possibility does not exist or is negligible, the person should not be considered as “identifiable”, and the information would not be considered as “personal data”.

The criterion of “all the means likely reasonably to be used either by the controller or by any other person" should in particular take into account all the factors at stake. The cost of conducting identification is one factor, but not the only one. The intended purpose, the way the processing is structured, the advantage expected by the controller, the interests at stake for the individuals, as well as the risk of organisational dysfunctions (e.g. breaches of confidentiality duties) and technical failures should all be taken into account.

On the other hand, this test is a dynamic one and should consider the state of the art in technology at the time of the processing and the possibilities for development during the period for which the data will be processed. Identification may not be possible today with all the means likely reasonably to be used today. If the data are intended to be stored for one month, identification may not be anticipated to be possible during the "lifetime" of the information, and they should not be considered as personal data. However, it they are intended to be kept for 10 years, the controller should consider the possibility of identification that may occur also in the ninth year of their lifetime, and which may make them personal data at that moment.

The system should be able to adapt to these developments as they happen, and to incorporate then the appropriate technical and organisational measures in due course.

Pseudonymisation

An alternative to generating anonymous data is to incorporate a mechanism to link identifiable individuals in one or more data sources to a de-identified record for that individual. Such linkage can allow the de-identified record to be updated periodically when the new health data has been acquired by the original data source, or for supplementary data about the same individual to be added from another data source. In such cases the de-identified data are supplemented by a “key-code", a unique identifier that can be generated, or mapped to, by any of the contributing data sources so that their periodic supplements of data can always be matched to the correct de-identified record. This key-code is known as a pseudonym, and the de-identified dataset is called pseudonymous rather

[8] This Working Party has been dissolved and new European guidance is expected from the European Data Protection Board. However, GDPR includes the wording “means reasonably likely” so the essence of this WP guidance still holds.

than anonymous. The action of generating keys and of retaining the connection between real identifiers and pseudonyms may sometimes be performed by a trusted third party on behalf of the data controller and research user. The GDPR considers pseudonymous data usually to be personal data.

Where a research user has no means of reversing the connection from a pseudonym to personal identifiers, the data may be considered anonymous from the perspective of that research user, provided that they have implemented appropriate safeguards and controls which restrict the possibility that an individual person can be identified. However, this scenario is open to interpretation within the GDPR and future national legislation across European Member States may differ in what extents of separation between a pseudonym and the true subject identifiers render the data effectively anonymous from a data protection perspective, or appropriately safeguarded. Please see:

- the extract, below, from an Opinion produced by the Article 29 Data Protection Working Party established by the European Commission to draft the European General Data Protection Regulation
- a publication by Mourby M et al. Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK. Computer Law and Security Review 34 (2018) 222-233. https://doi.org/10.1016/j.clsr.2018.01.002

Article 29 Data Protection Working Party

Opinion 4/2007 on the concept of personal data

Example No. 13: pharmaceutical research data

Hospitals or individual physicians transfer data from medical records of their patients to a company for the purposes of medical research. No names of the patients are used but only serial numbers attributed randomly to each clinical case, in order to ensure coherence and to avoid confusion with information on different patients. The names of patients stay exclusively in possession of the respective doctors bound by medical secrecy. The data do not contain any additional information which make identification of the patients possible by combining it. In addition, all other measures have been taken to prevent the data subjects from being identified or becoming identifiable, be it legal, technical or organizational. Under these circumstances, a Data Protection Authority may consider that no means are present in the processing performed by the pharmaceutical company, which make it likely reasonably to be used to identify the data subjects.

Appendix 3 Contributors to this report

The following members have contributed to the content of this report, participating in think tank meetings or by reviewing draft versions.

Editor
Professor Dipak Kalra, CHC Director of External Relations

CHC Co-ordinating Centre
Professor John Ainsworth, CHC Director
Clare Sanderson, Head of Information Governances
Gary Leeming, Chief Technical Officer
Mary Tully, Head of Public Engagement
Claire Smith, Head of Operations, CHC Hub
Carol Ann Costello, Business Relationship Manager

CHC NWC
Liz Mear, Director of NWC CHC
Julia Reynolds, Associate Director of CHC NWC
Debbie Parkinson, Public engagement lead, Innovation Agency
Professor Sumi Halal, Chair in Digital Health, Lancaster University
Dr Wesley Hutchinson, Academic GP, Lancaster University
Nick King Commercial Lead, CHC NWC
Victoria Neumann - Researcher
Jo Knight - Researcher

CHC Greater Manchester
Emily Griffiths, Head of Information, GM CHC
Zoher Kapacee, Head of Operations, GM CHC
Jo Hobbs, Public engagement lead

CHC Connected Yorkshire
Owen Johnson, University of Leeds
Kuldeep Sohal, Programme Manager, Connected Yorkshire

CHC NENC
Joe McDonald, Director of CHC NENC
Madeleine Murtagh, Professor of Sociology and Bioethics, Newcastle
Nick Booth CIO, CHC NENC

Patient representatives

Rob Ankers

John Massey

Industry representative

Dr Ed Conley, Chief Scientific Officer, AIMES

James Smedley, AIMES