A guide to data linkage

Contents

Click on any item in the contents list to navigate directly to that section in the document.

This guide to data linkage was produced to support NHS systems (such as ICSs and STPs), commissioners, providers, NHS England and NHS Improvement and arm’s-length bodies to co-learn and co-develop, and to share and spread best practice, learning and tools relating to data linkage that extends beyond primary and secondary care.

The guide was written in collaboration with experts from the following organisations:

- Alder Hey Children’s NHS Foundation Trust
- Care City Innovation CIC
- Connected Health Cities
- Health Economics Unit
- Kent City Council
- Maidstone and Tunbridge Wells NHS Trust
- Midlands and Lancashire CSU
- NHS Arden and Greater East Midlands CSU
- NHS Devon CCG
- NHS England and NHS Improvement
- NHS X
- SAS
- The Health Foundation
- University of Leeds

Aims of this guide

This guide is intended to provide information about the linkage of datasets beyond primary and secondary care. It has been developed for anyone involved in the planning and delivery of health and care, from GP practices and secondary care to integrated care systems. It aims to achieve the following:

- Introduce readers to the basics of data linkage
- Explain the benefits of data linkage
- Discuss when data linkage should and should not be used
- Set out best practice for creating and working with data linkage
- Outline key considerations and challenges
- Note pitfalls and issues to avoid
- Provide examples of data linkage in action

"This guide was commissioned to further help colleagues in the NHS, social care, community care, public health, local authorities and beyond to better understand the value of linking data, learn from those that have previously linked their data and see the 'art of the possible'."

Ming Tang – Chief Data and Analytics Officer (Interim)

"Data linkage is a critical step to better understanding our health economies and populations. Moving away from just healthcare activity data to a broader understanding of our population's health means we can make more informed decisions and better address the wider determinants of health."

Andi Orlowski – Director of the Health Economics Unit

Definitions and abbreviations

AI	Artificial intelligence
CSU	Commissioning Support Unit
DPIA	Data protection impact assessment
DSA	Data sharing agreement
GDPR	General Data Protection Regulation
ICS	Integrated care system
IG	Information governance
ISA	Information sharing agreement
JDCA	Joint data controller agreement
MPI	Master patient index
PHE	Public Health England
PHM	Population health management
PPI	Patient and public involvement
SNOMED	Systematised Nomenclature of Medicine – Clinical Terms. A standardised, multilingual vocabulary of clinical terminology used to exchange clinical health information
SUS	Secondary Uses Service – the single, comprehensive repository for healthcare data in England
UPRN	Unique property reference number

1. Background

This guide is intended to outline how to effectively deliver data linkage within health economies in England. It informs senior managers and system leaders about the reasons to link data from a range of partners and provides a technical roadmap for health economists, operation researchers and wider data analytics communities.

The scope goes beyond linking just primary and secondary healthcare data, emphasising the importance of also integrating local administrative data concerning wider health determinants and broader data integration across an integrated care system (ICS). It also provides expert comment on best practice in data linkage and highlights potential risks and mitigations.

This document is intended to be an informative guide describing the key steps in creating and using data linkage that is:

- Practical
- Evidence based
- Freely available within the NHS
- Transparent in methodology and processes

It has been written in modular sections and does not need to be read in order from start to finish.

To inform this guide, interviews were held with 11 subject matter experts to provide context and initial data about experiences with data linkage, obtain expert guidance on leading approaches to data linkage, and identify topics to discuss in more detail.

A series of six workshops was then held to provide a forum for the same subject matter experts to come together with analysts to discuss issues raised in the interviews.

The quotes included throughout this report are taken from the interviews and workshops.

The ‘how to’ section (section 5) includes a practical, stepwise guide on how to deliver data linkage, and the ‘data linkage in action’ section (section 7) includes real-life examples of how some health systems have scoped, delivered and used data linkage.

2. What is data linkage?

Data linking means bringing together two or more sources of information which relate to the same individual, event, institution or place. By combining the information it may be possible to identify relationships between factors which are not evident from the single sources.[1]

Within the health sector, data linkage is commonly understood to refer to connecting health data beyond primary and secondary care, and perhaps also incorporating social care datasets, delivering a real-time picture of how population health and healthcare demand is changing over time. Incorporating wider determinants of health enables users to look at the whole person, rather than focusing solely on aspects of health and social care.

Linking datasets together should be viewed as an investment in a cumulative store of knowledge which can be used to improve the health of the whole population.

To link data from multiple sources, common identifiers must be used, such as NHS numbers or unique property reference numbers (UPRNs), which are unique numeric identifiers for every spatial address in Great Britain.

The NHS uses the NHS number, but local government and wider public sector datasets do not have a single identifier (national pupil reference numbers and National Insurance numbers are unique identifiers but are only given at certain ages; they are not cradle to grave). As a result, data is often linked using the NHS number to create a master patient index (MPI). As long as all

[1] https://cms.wellcome.org/sites/default/files/enabling-data-linkage-to-maximise-value-of-public-health-research-data-phrdf-mar15.pdf

datasets which are to be linked contain the NHS number, information on services, personal particulars, clinical mobility, socioeconomics and more can be integrated into a single linked dataset.

> *"There's a lot of work to do to not just make it health and social care but also include some of the other public health datasets. There is definitely value in starting to look at the person as a whole, because none of these factors [directly] affect their health, and if you have to do some proactive management, it helps us understand all that."*

The MPI provides longitudinal data that is regularly updated with details about how many people are born, have migrated into or out of the population, or have died, giving rise to a rich, deep and broad dataset which can then be used to derive significant value in a range of contexts and use cases.

The ultimate benefit of a linked dataset is that it can allow for tracking of depersonalised data of each and every individual from the moment they were born – personal particulars, clinical morbidities they have developed over time, what services they have utilised at what point – all the way up to death.

That continuum of information and activity, ascribed at each individual level, forms the basis of population health, because it supports aggregation of that information from individuals through to a cohort and from there to the whole population.

Once data linkage has been achieved, it is possible to segment the data to support population health management (PHM; a technique for using data to design new models of proactive care and deliver improvements in health and wellbeing which make best use of collective resources to improve physical and mental health outcomes, promote wellbeing and reduce health inequalities across an entire population) using tools such as IMD, MOSAIC and ACORN and real-time administrative data collected by local authorities.

3. Why use data linkage?

Data linkage offers a more comprehensive understanding about the health needs of a population than can be gained from any single dataset, offering the opportunity to support and inform decision-making by health systems.

Current standalone datasets and healthcare indicators are useful for descriptive analytics such as performance management and surveillance and PHM but not for diagnostic, predictive and prescriptive analytics such as evaluation, forward capacity planning and impactibility modelling, which aims to identify patients for whom preventive care is expected to be successful.[2]

Going through the process of linking data can help identify the strengths of each dataset in terms of specific variables (e.g. ethnicity) and provide a useful focus and oversight of where there are weaker links in understanding health and care.

Using data from organisations other than primary and secondary care (e.g. local authorities, social care and community care) can help local systems such as ICSs understand vulnerable groups in their areas whose health needs are not being fully addressed; for example, people on a housing waiting list, those with drug or alcohol issues, people with learning disabilities or autism, traveller communities, veterans and those affected by domestic violence. Insights from data linkage across systems therefore have the potential to identify unmet needs and highlight opportunities to support vulnerable individuals.

[2] Lewis GH. "Impactibility models": identifying the subgroup of high-risk patients most amenable to hospital-avoidance programs. Milbank Q. 2010;88(2):240-255. doi:10.1111/j.1468-0009.2010.00597.x

If possible, linked data should be integrated at either the person level or, where applicable, at a household level, as many health-related factors are also household related.

3.1. Turning data into insight

Data alone is simply a collection of facts such as numbers or characters. Without context, data can mean very little. For example, 12012012 is just a sequence of numbers without apparent importance; but if we have the context of ‘this is a date’, we can easily recognise it as being 12 January 2012. By adding context and value to the numbers, they have more meaning.

In this way, data becomes information, which can be used to gain knowledge, which can be interpreted for wisdom. This approach is shown through the data, information, knowledge, wisdom (DIKW) pyramid (Figure 1), which represents the relationships between each level. Each step answers different questions about the initial data and adds value to it. The more we enrich our data with meaning and context, the more knowledge and insights we get out of it so we can take better, more informed, data-based decisions.[3]

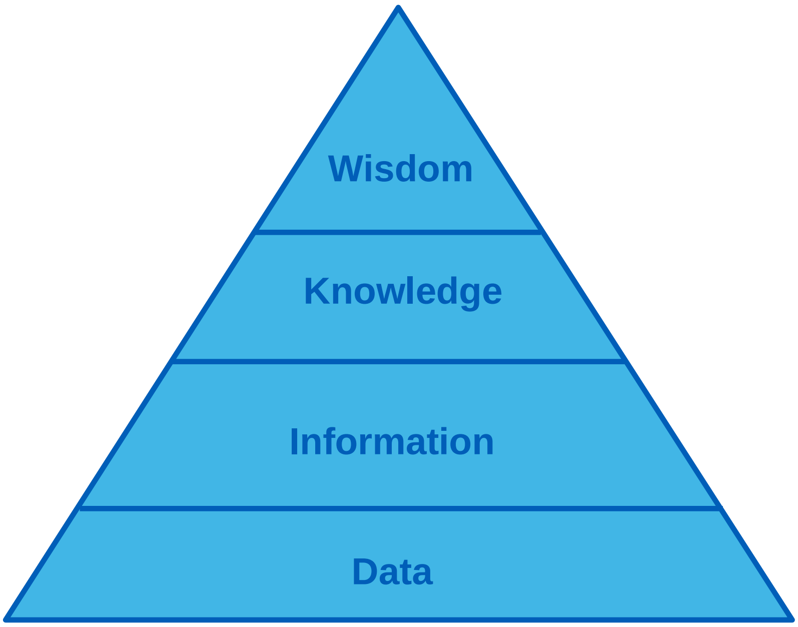


Figure 1: The DIKW pyramid

A good example of how data linkage can help service users is through the better understanding of the impact of long-term conditions and multi-morbidity on the population, which represent the majority of healthcare spending.[4,5] Its application within PHM to aid understanding of the impact of interventions can also lead to changes which ultimately improve outcomes and benefit patients and service users. (This data linking guide describes the value of data linking for PHM and the journey that systems need to follow for an agreed cross-system approach to linking data, and contains case studies on systems that have already done this successfully.)

Data has been collected for many years on the prevalence of single long-term conditions in the population (e.g. diabetes, heart disease, cardiovascular disease, stroke, chronic obstructive pulmonary disease, asthma and dementia) as part of the quality outcomes framework to build the nationally correlated anonymised registers. This data shows the proportion of people with a long-term condition in each financial year and can be used from a practice through to a national level. However, it does not show what proportion of those patients have more than one long-term condition.

[3] www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/

[4] www.england.nhs.uk/ourwork/clinical-policy/ltc/house-of-care/

[5] www.kingsfund.org.uk/projects/time-think-differently/trends-disease-and-disability-long-term-conditions-multi-morbidity

As well as giving a better understanding of the impact of multi-morbidity on a population's health, linking longitudinal data from multiple data sources can enable retrospective analysis incorporating case–control techniques.[6]

It also supports checking and improvement of data quality. All data contains errors to a greater or lesser degree; combining multiple datasets allows the consistency of data to be checked and may enable the gaps to be filled in. Real-life examples include[7]:

- Linking data from midwives to vital events registers showed that previous estimates of births in one ethnic group had been misclassified to the dominant ethnic group
- Researchers at the Karolinska Institute demonstrated that the use of linked microdata reversed the findings from area-level statistics about the impact of a GP engagement programme
- An Australian study linking multiple cancer registries showed that the 'official' register was underestimating cancer incidence by about 12%, largely due to non-standardised variable management

Pooling data from different years and data sources (perhaps even different countries) can also generate sufficient data to be useful in analysing and modelling rare events. By their nature, it is difficult to generate sufficient information on rare events from single data sources. For example, Marshall–Smith syndrome has approximately 23 sufferers worldwide; without data sharing, no effective analysis is possible.[6]

Data linkage can provide a more visible understanding and greater clarity of patient flow through the system on both an individual and a population basis. The real-time data enables users to look at resource allocation with a much more informed eye; it can highlight where people are falling through the gaps or presenting many times, and link out-of-hours data with data from acute and wider health and care providers. Using linked datasets can enable better understanding of relationships between factors that may not otherwise be easily connected. For example, since the Covid-19 pandemic there has been analysis of people who are not using services or those who are on multiple waiting lists and who therefore default to attending A&E; this kind of analysis is only possible with a linked dataset.

3.2. Moving from descriptive to prescriptive analytics

Data linkage is essential for forward planning and enabling prescriptive analytics to support impactibility modelling, which describes the subpopulations that will benefit from a range of interventions. Capacity planning is often based on linear trends and/or broad assumptions; it should instead take into account a number of complex interrelationships. For example, if working to predict the number of beds needed, decision-makers must consider many interdependencies and interrelationships between socioeconomic risk factors, demographic drivers and healthcare utilisation activity. Changes in the way services are delivered (e.g. new clinical pathways) or changes in activity will also affect bed supply over time. Data linkage can give a wider real-time picture of the population's health needs which can help mitigate this complexity.

It is important to embrace all aspects of analysis and not fall into the trap of thinking that some are superior to or more advanced than others. Sometimes a high-quality descriptive analysis is most useful; sometimes the question is best addressed through a form of predictive analysis. Whatever type of analysis is most appropriate will be determined by the question being addressed – getting that right is a fundamental analytical skill – and its value will be determined by whether the decision-making system is designed to draw on the analysis. A recent publication by The Strategy Unit '[Advancing the analytical capability of the NHS and its ICS partners](#)' unpacks the components

[6] One of the drawbacks of using a longitudinal approach to investigate the causes of disease with low incidence is that large and lengthy studies may be required to give adequate statistical power. An alternative is case–control (or case–referent) design. In a case–control study, patients who have developed a disease are identified and their past exposure to suspected aetiological factors is compared with that of controls or referents who do not have the disease. www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated/8-case-control-and-cross-sectional

[7] https://cms.wellcome.org/sites/default/files/enabling-data-linkage-to-maximise-value-of-public-health-research-data-phrdf-mar15.pdf

of analysis and provides a useful basis for systems/organisations to assess whether they have the right mix of capabilities and capacity.

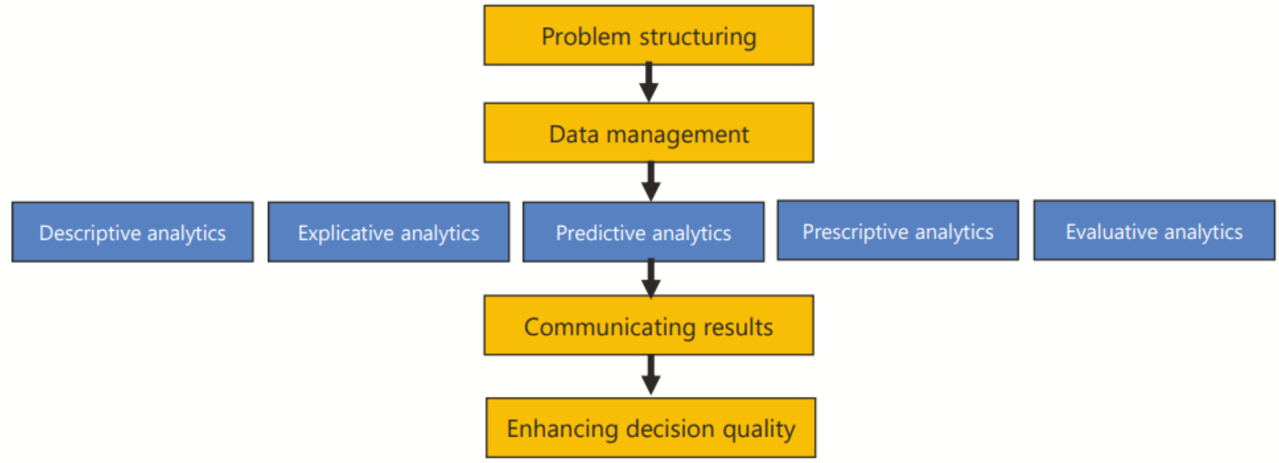


Figure 2: Analytical projects typology[8]

3.3. Weaknesses of silo data

Most existing datasets and indicators, when used in isolation, have some weaknesses. Individual datasets are often aggregated and/or spatial based (e.g. ward-level/LSOA-level statistics, MOSAIC, CACI and ACORN), which masks variations in characteristics at a person level.

Some datasets need updating regularly, while others have more longevity. However, most are not updated in real time (e.g. PHE Fingertips is only updated annually or biannually) and they are not always useful for analysis of local population health needs, particularly when evaluating change or the impact of a specific intervention in real time.

By linking data across different datasets, many of these weaknesses can be addressed. Siloed datasets can be useful for reporting, but the greater depth of data in linked datasets means they can also be used for analysis and planning.

In addition, analysts carry out numerous analyses to correlate high-level anonymised indicators such as deprivation and emergency admissions. Data linkage supports a move away from correlation analysis – which involves comparing two different indicators that are not linked together – towards regression analysis, which can be used to understand the adjusted effect of different risk factors and how they vary in line with each other.

Regression analysis can be used to identify which variables have impact on an issue such as emergency admissions, and the relationship between those variables. Linked data allows regression analysis to be more reliable as fewer assumptions are being made, and so the reporting can be significantly deeper.

3.4. Completeness and quality of data

Linking datasets can help to identify where an individual dataset may be incomplete or inaccurate and allow this learning to be fed back to organisations. Looking at linked data may highlight, for example, where NHS number coverage is low in social care, or where ethnicity is recorded differently in different datasets.

Linkage also helps to overcome such data quality issues in individual datasets, as some datasets

[8] Extracted from ‘Advancing the analytical capability of the NHS and its ICS partners’ produced by The Strategy Unit. Available at *https://www.strategyunitwm.nhs.uk/publications/advancing-analytical-capability-nhs-and-its-ics-partners*. Reproduced with permission.

are more accurate and complete than others. The combined sum of all the linked datasets is therefore more complete and accurate than any individual dataset.

In many cases, knowing that data is being shared drives individuals or organisations to improve the quality of the data they create or curate as they better understand its wider uses.

3.5. A range of insights

Data linkage can be used to support work in modelling and understanding service improvements. With appropriate information governance (IG) agreements in place, it can be a hugely valuable resource for research projects and can provide valuable insights into factors or behaviours that drive better health. The range of data included can mean that new bespoke data collections are not needed for every new project. Many integrated datasets already have primary care, hospital, local authority and social care, community care and mental health data that will support a wide range of system-wide analysis requirements.

Data linkage across a range of organisations can enable system collaboration and understanding of the whole system and whole population through advanced analytics, and provide better estimation of the attributable effect of different factors.

3.6. Improved understanding of relationships

Data linkage underpins integrated care and PHM, allowing organisations to break out of their silos by joining different data. This allows longitudinal records across settings to be evaluated; the wider the variety of data available, the higher the degree of understanding of any relationships.

More information about systems thinking, which relates to organisational interdependencies, is available here (competencies K5, S3 and S6).

Linked cross-sector data provides the opportunity to develop innovative products or tools to better understand the drivers behind wider determinants. For example, women’s life expectancy stalls in deprived areas; how does that manifest in air quality/housing/spending data? How might that data be used collectively to address those issues?

4. When not to use data linkage

There may be times when the data linkage approach is not appropriate, often when there are concerns around the privacy and confidentiality of individuals, the purposes for which the data is to be used (e.g. where this may be considered unethical) and/or the intended recipients of data. It is important to stay up to date with changes in the law (e.g. changes to the application of the GDPR as a result of Brexit) and best practice guidance such as the Secondary Uses Data Governance Tool (subject to testing and further development) and to ensure that all partners sharing data have the required systems in place to guarantee the correct procedures will be followed and monitored.

Ethical and safety risks can include key elements such as a senior officer not being in place or the lack of a fully expressed, clear purpose for linking data.

There may also be practical concerns around weak linkage (due to differing levels of data quality) or incomplete datasets. Both of these could lead to misleading conclusions and increase health inequalities. For example, when dealing with rare disease data, linkages with cohorts that are not statistically significant could have unintended consequences.

> *“Ask the right questions in the system, understand what analytics methods are needed to answer them, then determine whether data linkage is required. If you’re just measuring indications such as life expectancy, it’s not needed.”*

5. How to use data linkage

A step-by-step guide to the technical aspects of how to carry out data linkage is included in the appendix to this guide. This 'how to' section focuses on practical aspects of setting up and optimising a data linkage project.

Note that it may not be necessary to create a new linked dataset for your project. Before starting any data linkage work, first check what may already be available in your area; your local commissioning support unit (CSU) may be able to offer support.

5.1. Questions and analytics

Before any data linkage work begins, it is important to agree the questions that the linkage is designed to answer and the type of analytics needed to answer those questions. Forward planning, simulation modelling and prescriptive analytics necessitate different sets of analytical methods. Once the question and analytical approaches are understood, then decisions can be taken about what dataset needs to be created.

5.2. Senior leadership

5.2.1. Understanding

Senior leadership, particularly chief executives of local organisations, need to have a clear understanding of data maturity and be open and honest about the need for a linked dataset, in order to understand the value and importance of working together. It is helpful to have senior leadership champions with a good understanding of data analytics or a background in informatics (e.g. a chief analytical officer) who can drive the conversation and encourage more investment in developing the linked dataset. Senior managers need to understand the potential impact of data linkage. Using case studies and linking in with teams that already have linked data in place will help here.

5.2.2. Skills

It is crucial for senior leadership to be equipped with the skills for setting out the complex questions to frame or the strategic priorities to address. This avoids the risk of limiting the scope of the data linkage work to operational business intelligence and surveillance, when it is better suited to evaluation for planning and broader exploratory research.

5.2.3. Support

Support for the data linkage project from senior managers and clinicians is critical in order to follow through on insights that come from the analysis. Once the evidence base for action exists, there needs to be authority to support action. Make sure you get trust and buy-in from all the organisations involved. Even if they do not understand the 'how', the 'why' is important. Help them understand the way the data will be interpreted, and articulate the benefits: that information gets fed down to clinicians and others inputting the data and leads to improved data quality from the beginning. Get support for this from experts in the organisations that are sharing the data (e.g. mental health analysts know mental health data best).

5.3. Information governance

Having a data sharing agreement (DSA) in place with the data controllers and/or NHS Digital is essential. Be aware that establishing that agreement can be a lengthy process. The work not only involves putting a document in place and getting it signed; it is also about putting in place a meaningful and effective governance framework which enables all the data controllers who are accountable for the data being shared and linked to meet their legal requirements. This may include setting minimum standards which partner organisations must meet in relation to data protection and confidentiality compliance, security and data quality. It should also empower data controllers to determine what data is shared and linked, the purposes for which it is used (and not used) and the recipients of that data. The DSA should clearly describe the various roles and responsibilities and the mechanisms which have been established to ensure everyone is able to comply with the law and clearly demonstrate that compliance.

Remember that there may be sensitivities around stigmatised data such as HIV status and that linking records in certain settings could lead to potential abuse. If you are considering working with data which could be deemed sensitive, ensure a robust data protection impact assessment (DPIA) is in place as well as a full reporting and management strategy to ensure no leakages of personal information can arise.

Under current data protection legislation, all health data is deemed to be ‘special category data’ and treated equally. However, the Royal College of GPs has defined a ‘sensitive dataset’ which includes data such as sexual health, HIV, fertility treatments, pregnancy terminations and gender reassignment.

It is also advisable to consider what the public expectation is likely to be about the ways in which data may be used. Public perception of exactly what is ‘reasonable’ changes over time and should be taken into account.

See section 6.1 below for more detail about information governance.

5.4. Data

5.4.1. Quality

Data quality is key. Organisations have different capabilities and capacities in terms of how they enter data on their systems and then use the data for direct care. It is not unusual for clinicians to use different approaches across the system; within primary care, GP practices may not all code the same way. It is important to take time to understand the baseline and take account of these different coding approaches before any conclusions are drawn regarding outcomes.

Make sure you understand the relationships between separate elements of data – units of measure and the timings. For example, the patient record is descriptive about the patient, but hospital data is measured in a very episodic, activity-based way.

5.4.2. Standards and consistency

Standardisation and consistency across organisations are important for a data linkage project to succeed. To support this, look for a structured data specification such as PRSB standards, source data from systems which are based on a structured specification wherever possible, and encourage organisations to implement the common standards if they are adopting new systems.

While it may seem a minor detail, consider the format in which NHS numbers are recorded in different systems, as variation here can pose technical challenges in linking data from different systems, even when a common identifier is held.

There should also be an agreed solution to the challenge of working with different units of measurement across different datasets (e.g. size/weight, prescribing), as this can cause issues when analysing at scale.

At the beginning of a data linkage project, look for elements of consistency across datasets – if one is all males and another is all females, that should raise a red flag. Make sure data consistency checks are in place on receiving/processing the data extracts throughout the project (consistency of volume, completeness, time periods etc. should all be checked).

Decide from the start which dataset is the master record (it should be the one you have the most confidence and trust in) and link other datasets to that one. It is a good idea to document the provenance of data so that others can clearly see where each element originated.

Consider where data conflicts may arise, and establish a process for resolving them where the same data item for the same person is different in different datasets. For example, addresses, dates of birth and ages can be recorded differently across different providers, and care providers may be registered in different addresses/postcodes if there are many sites or previous mergers.

The key in all aspects of data management and standardisation is not only to establish a clear process, but also to clearly show how that process has been implemented in the end dataset.

Be aware that not everyone uses NHS numbers, and this can affect the quality of the data linkage. Assuming that name, date of birth and postcode are collected or held, this can be addressed, but it should be considered at the outset. Other identifiers (e.g. the national pupil reference number) may be available for use for certain cohorts, but IG, data protection and any potential data sensitivity should be considered at all stages.

It may be appropriate to develop regional data quality protocols with agreed standards in line with national standards for key items. For example, the address data standard BS7666 could be adopted to ensure consistent spacing in postcode data, or a product such as the Ordnance Survey's AddressBase®, which matches 29 million Royal Mail postal addresses to UPRNs, could be useful.

The approach to opt-outs and flags should be checked in all datasets to ensure consistency throughout. Ideally, data relating to any individual who has opted out of having their data shared via the national system[9] should be removed before the data is submitted. Alternatively, if IG requirements and the relationship between data controller and data processor allow, the data can be flowed into a central repository first and then have the opt-outs removed. In both cases, the best option will be to use the national system. The only exception to this is data held by GP surgeries, as current NHS policy requires locally recorded opt-outs (Type 1 opt-outs) to be applied at source.[10]

Creating a standard approach to flags can be more difficult, as certain flags can be coded differently at the local level. The team in Kent (see Kent Integrated Dataset) addressed this by agreeing a centralised methodology, creating all flags centrally and ignoring any local flags.

> *"With primary care we follow a template-based approach and ask them to follow standardised coding. Secondary care has more national guidance because they have to submit nationally; primary care has a lot more freedom."*

5.4.3. Profiling and bias

It is commonly understood that certain populations are less likely to provide certain elements of data, and as a result are likely to be missing from or underrepresented in a dataset. This phenomenon can lead to unintentional bias in the resulting analysis and should therefore be considered at the start of any data linkage work.

> *"We struggle with collecting ethnic origin data as it's not regarded as very accurate. I don't know that we've got the data to properly look for those kinds of biases."*

Look at diversity and inclusion within the team that is undertaking the analysis and consider whether you are fully cognisant of any possible unconscious biases or lack of knowledge that might impact the results.

There may be bias in terms of the way people are grouped, or an individual might have declared potentially sensitive information such as sexual orientation, ethnicity or homeless status to one

[9] https://digital.nhs.uk/services/national-data-opt-out

[10] www.nhs.uk/using-the-nhs/about-the-nhs/opt-out-of-sharing-your-health-records/

NHS organisation but not to another. This raises the question of whether it is appropriate to pull the declared information through to a linked dataset.

> *"I always use deprivation index as a characteristic of any data linkage, because it's a proxy for so many other things."*

Think about the population the resulting analysis will affect, and what the impact of that analysis might be on every constituent part of that population. Check whether you have a good understanding of how different communities and populations are represented within the data and consider how to involve the people whose data you are using and the people who will be impacted by it. This will also help support data protection compliance by ensuring you are being transparent with data subjects about the way their data is being used and shared.

> *"Hard-to-reach groups who are not engaged are probably the people who need more help."*

5.4.4. Completeness

Data completeness is often an issue, with some datasets having gaps when data stopped flowing. Take a systems perspective and consult widely with system leaders and patients about what datasets could be missing from the work (e.g. links to data in the charity sector) or whether any of the datasets could have experienced interruptions to data flow or may be otherwise incomplete.

Knowing what data is missing is as important to report as what data is present for robust analysis. For example, percentage figures are often given for the matched patient population excluding the unknown patient population, which may exaggerate the strength of the observation.

Carry out exploratory data analysis to identify fields with more missing values than others and try to find out how completely each field should be populated. Liaison with data suppliers will be key to an understanding here.

Consider how material each field might be to the analysis. Certain fields should always be populated, but others may not always apply. Where unknown or null values are present in key fields, work with organisations to get them populated.

Certain sections of the population are more likely to be less well coded, so certain data elements may be underrepresented in the dataset. Consider how this will impact on analysis for care planning, for example.

5.4.5. Population size

There is no optimum population size for data linkage, but it is important to consider whether the records within the dataset reflect the make-up of the population in terms of factors such as ethnic distribution, age, gender and long-term conditions.

Generally, the larger the dataset, the more robust the analysis of outcomes will be. Datasets that are too small may have insufficient variance in socioeconomic factors or insufficient overall data in all disease areas to analyse.

Consider that bigger does not necessarily mean better; biases which are more obvious in smaller linkages can be hidden in large ones. Using a bigger population is also likely to mean involving more organisations, which can lead to challenges in managing relationships and infrastructure between and within organisations.

This means there is a maximum population size that works effectively, particularly if multiple datasets are being joined. Systems should consider linkage of smaller datasets (possibly at primary care network level) for local interventions, and larger projects for place-based or strategic analysis.

Another factor for consideration is how to draw the boundaries, taking into account travel distances and commuter patterns. There are also difficulties in areas such as Cheshire, because joining records between England and Wales is much more difficult.

5.4.6. Matching techniques

Data matching in linked data is the process of identifying and merging duplicate data records, for example records from health and local authority. The aim is to de-duplicate and clean records as needed to create the most accurate and representative dataset possible, which can be used as the patient master list.

There are a number of issues that can be raised at this point; for example, when individuals give different information to different organisations, or log address or name changes with only some organisations. It is vital to understand when apparently separate records actually relate to a single individual.

Different data matching techniques will help provide a clear view of the relationships between different records despite potential differences in recording methods and data quality. The two main types of matching methods are *probabilistic* and *deterministic*.

Deterministic matching uses an exact match on a unique identifier to merge multiple records. In the health field this will usually be the NHS number, but if the NHS number is not included in some records (e.g. local authority records), address details can also be used. Additional identifiers that may be useful include phone numbers and email addresses.

Probabilistic matching (also known as ‘fuzzy matching’) uses algorithms to score and weight the variables and inconsistencies present in records to determine whether the records held by different organisations belong to the same person. This will help determine whether individual records should be linked, depending on whether they reach a certain threshold.

Techniques such as probabilistic matching (in which homophones, transpositions and other non-exact matches are used)[11] can give rise to concern in some situations. A checked NHS number should be used wherever possible, but where this is not available, probabilistic matching – though not perfect – is a valid linking methodology.

In practice, many organisations do not have an NHS number for some patients, so some kind of probabilistic link or lookup is necessary; however, it should be used sparingly. Inferences based on a linkage that results from probabilistic matching could miss a significant portion of the population, and the resulting outcomes should not be taken as ‘gospel’.

> *“It’s OK to use in aggregate, but at a personal level, it could mislead.”*

Probabilistic matching is often useful and acceptable when looking at bigger-picture issues, but you should be aware that there are potentially significant risks associated with using it at a more individual level, and caution is advised when making assumptions about probabilistic matching if looking at smaller populations or those that are not well represented in the dataset as a whole.

However, as long as the risks of probabilistic matching are understood and mitigated, the process can allow important data to be linked that can give useful insights. A monitored data quality reporting approach used alongside probabilistic matching can help to mitigate the risks.

> *“With fuzzy matching, there is a risk of making decisions based on wrong or incomplete data.”*

[11] https://digital.nhs.uk/developer/api-catalogue/personal-demographics-service-fhir

Over time the algorithms used in probabilistic matching have improved, and it is now easier to spot a ‘false join’. However, users should be mindful that when probabilistic matching is used, it is not possible to reverse-engineer data and use the dataset to reidentify patients at risk.

5.4.7. Quality assurance

Some organisations which have successfully used data linkage have developed a quality assurance approach using data points such as accuracy, completeness, timeliness and auditability by, for example, requiring an operational or clinical lead to confirm that datasets ‘look right’.

> *“Ask someone who uses the data to talk you through it and sense check – they might say ‘oh, that prescription changed years ago’ or ‘that should be centimetres, not millimetres’.”*

Data can be de-personalised and made available to analysts in care providers to enable them to compare it with their local system and flag any discrepancies.

Compare trends over time to check data quality; if a figure is relatively consistent and then suddenly changes, investigate the cause.

Artificial intelligence (AI) methods are available which could be explored further to improve data quality and data matching.

5.4.8. Trusted research environments

The use of secure data environments, such as Trusted Research Environments (TREs), enables the highest standards of information governance, transparency and security by removing the need for data to be physically shared between users.

In TREs data remains within a secure environment, is analysed in situ and only accessed by those whose credentials have been established by an accredited authority. It is recommended that TRE providers offer data linkage services with controls on access that are proportionate to the approved use. As data assets in TREs are held in a safe setting, any export should be subject to appropriate data minimisation.

At the time of writing, NHSX planned to deliver the following in 2022:

1. A minimum technical specification for TREs – covering core areas such as interoperability, cybersecurity and use of privacy enhancing technologies.
2. New and enhanced TRE standards and policy – NHSX will also develop policy and best practice guidance for TREs covering, for example, when their use will be mandatory and any legitimate exceptions.
3. An accreditation framework – detailing the specifications and standards that TREs must adhere to, and how adherence will be assessed and monitored.

5.5. Other considerations

5.5.1. The team

Effective collaboration among the data linkage team is important. Consider setting up a steering group for participating organisations with clinical input, ensuring a proper governance process is in place, and having one lead team in one organisation to store, anonymise and manage the data.

Have a data linkage champion in each part of the system. Involve experts in each included dataset; they know their data best.

Identify the people who will do the extraction processes and consult with them early on – remember this will have a big impact on their workload.

5.5.2. Commercial sensitivity

Be aware that there may be some commercial sensitivity around data being shared between competing providers. This can be addressed through robust governance.

5.5.3. Technical infrastructure

Data linkage work needs robust technical infrastructure and a high-performance environment; make sure you have the necessary infrastructure to deliver the project. Involving IT staff with the project from an early stage can help to properly identify the requirements and timescales and assist with the secure networking demands that are often required when gathering data across disparate systems and services.

The ‘What Good Looks Like’ framework produced by NHSX describes how arrangements across a whole ICS, including all its constituent organisations, can support success. There is an expectation that the standards within the framework will be used to accelerate digital and data transformation at both a system and organisation level.

5.5.4. The timeline

Healthcare is a fast-moving landscape in which GP practices merge or close and organisational boundaries and structures change. In order to carry out longer-term analysis, data relating to patients who are deregistered during the time period being analysed should be retained within the dataset, but the ability to remove them from the cohort should also be retained in order to maintain analytical robustness and avoid bias in the results.

5.5.5. Publishing protocols

Ensure your approach to data linkage can be shared locally and nationally by publishing protocols and considering ahead of time what to get involved in and where the outputs from the project will appear (e.g. conferences or publications).

5.5.6. Ongoing improvements

Consideration should be given to creating an environment that is extendable and which can be linked to new datasets that may be available in future, and to how this could be done. Having a well-considered and repeatable methodology for adding new data will allow you to respond both to the changing healthcare environment and to requests for new linkages to add insights should they become available at a later date.

5.5.7. User engagement

Support should be available to users once they are able to access the system you have produced so that consistent, correct information can be delivered. Part of this can be through adding ‘fact’ information to your dataset to aid with consistent, documented calculation methodologies as part of your linked dataset. High-quality examples, user groups and shared/open code repositories can also help people use your dataset in a consistent way. User groups can offer an important feedback loop on any identified data quality issues.

6. Potential challenge areas

A number of challenges may be associated with using data linkage; users should be aware of these and prepare for them in advance. Many of the challenges are centred around IG and managing relationships and communication between different organisations; there may also be practical challenges relating to the technology needed to link or reidentify data securely, and to aspects of transparency such as consulting with the public about particular use cases for the data.

Questions also remain around where local government fits in. National NHS policies have been developed, advocating and emphasising the need for local government to take part in PHM; however, there has not yet been a significant policy shift from the local government side.

Overall project governance needs arranging before IG is confirmed, and a single set of processes and procedures should be established around authorising new data linkage and data access requests. Consideration should also be given to defining a clear purpose for the data linkage and ensuring any required secondary use permissions are in place.

The following sections go into further detail about some of the challenge areas. This section is not intended to be exhaustive, and users should be aware that in any specific project other challenges may arise.

6.1. Information governance

Experience has shown that a majority of the public do not realise their data is not already being shared within the wider health system. The main concerns many people have are around the security of their data, who it is shared with and transparency about the purposes for which it is used.[12]

IG designed to safeguard data linkage projects should be built in from the start. IG for data sharing and linkage can be a challenge, with potential variation between individual and organisational interpretations of IG rules and appetite for risk. Guidance from NHSX on shared care records and information sharing is available online here.

Securing the necessary level of IG clearance to access the required data at the level of detail needed from each organisation is often time-consuming and complicated. The greater the number of organisations involved, the more difficult the project is, with separate data access requests for each organisation. Work is under way centrally to simplify this, including an IG framework, data sharing framework and national data strategy. Conversations can, however, now be had in advance of central guidance becoming available to agree a common process for submitting and considering requests for access to linked data and devising common assessment criteria against which requests can be considered and then authorised or rejected in a consistent manner.

Other IG challenges include being limited in terms of only using data collected for administrative purposes and developing a pseudonymisation approach[13] that satisfies the need to link with common law duty of confidence. Currently, NHS Digital is the only organisation with the statutory power to link NHS and non-NHS data without encryption, although other organisations do have the legal basis to do so, subject to following correct procedures. More information about the legal basis for data linkage is available here.

The content of datasets can make data access more complex if they contain sensitive information or if individual-level data or data about small numbers of individuals can be accessed (the latter increases the risk that individuals may be identifiable even if other pseudonymisation or anonymisation techniques have been used). Significant resource is required for data governance, particularly in non-NHS, non-social care settings.

Data linkage requires a single process – for example, having a common template for a DPIA – and an experienced, senior group that represents the system (usually an ICS). That could include one nominated data protection officer, a Caldicott Guardian, clinician, clinical lead, research lead, business intelligence lead and so on. This group should create the governance framework and supporting processes that facilitate decisions and authorisations around access, new data requests, data access requests or linked data requests. The group should also include a patient and public involvement (PPI) lead supported by adequate capacity and funding, because data protection principles require that organisations demonstrate how they are engaging with the public around the use of their data.

[12] www.gov.uk/government/news/transparency-and-public-engagement-are-essential-to-demonstrating-that-data-is-being-used-for-public-benefit

[13] Pseudonymisation involves replacing names or other identifiers which are easily attributed to individuals with, for example, a reference number. The reference number can be linked back to the individual if the user has access to the relevant information. Anonymised data is stripped of sufficient elements that mean the individual can no longer be identified. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-is-personal-data/

It is important to complete new engagement and insight and testing with the public through existing PPI networks around secondary uses of data. However, organisations also need a more holistic, whole-system investment in PPI engagement focused on data usage.

6.1.1. Overcoming the challenge

Effective governance is crucial to the success of a data linkage project, so focus on it from the start. Put in place direct data control, an ethics committee and a governance framework across the board, and make sure you are acting lawfully and specifically in relation to common law duty of confidence. Remember that Caldicott approval may be needed, and engage Caldicott guardians early in the project.

DPIAs and DSAs can only be developed once the preparation steps above (who, what, when, where, how, why) have been sufficiently defined (with the input of IG subject matter experts who can provide helpful advice on what may be lawful and achievable).

The governance structure should not only manage data security but also address any potential commercial sensitivities linked to sharing data between, for example, competing provider organisations, or between a commissioning organisation and a provider which may later bid for a contract with the commissioner.

Tools are available to manage the IG process – for example, software to manage signing of an information sharing agreement (ISA), DSA or joint data controller agreement (JDCA). Explore the use of data institutions to link and broker access to novel datasets or datasets where there are issues around trust or governance.

To support relationships, build a good governance group that has the right executive buy-in – potentially as part of a wider analytical and intelligence function.

Using evidence from people who use a service related to the project can help to break down some of the barriers to data sharing. Case studies of how linked data is used to improve patient care can help people see this as less of a ‘data thing’ and more as part of a wider process. One interviewee explained that his team went to seven different organisations and asked 250 patients whether they would be happy for an NHS employee to use their data to provide a better service for them; 90% of those interviewed did not realise the data is not currently shared.

If an ICS invests into regular PPI, this can be used for testing ‘reasonable expectations’ around data integration; for example, linking domestic violence data from police. This PPI could be used to generate evidence of meaningful engagement and insight that would hold up in case of a judicial review around local processes and procedures.

6.2. Structure of the NHS

The disparate nature of the NHS’s structure is a related challenge. Many patients perceive the NHS to be one organisation and are surprised to learn that their data is not already being shared within the NHS. ICSs are expected to develop a system-wide intelligence function.

> *“If the NHS was one big company instead of a collection of separate organisations, it could simply share data across the company and much of our data linkage wouldn’t be seen as external data linkage; it would just be seen as using its own data assets to do its business.”*

Different organisations record data in different ways. This creates additional challenges to linking data if the organisations involved do not have a standard way of recording common data fields such as names, addresses, postcodes or dates of birth. Not all organisations can legally hold an NHS number, so as data linkage is expanded across systems a different set of matching criteria would need to be used, even with all the relevant agreements in place.

Senior leadership buy-in is crucial, because creating a linked dataset requires the support of all the organisations taking part and a willingness to share data. Organisations must work together, with terms of reference providing formal governance. With the advent of ICSs, there should now be increased willingness for NHS trusts and commissioning organisations to work together.

In some cases, data linkage is not seen as a priority; organisations see internal support as more important than system support, and an analyst in an acute hospital will often look only at the acute hospital and not beyond its walls.

6.2.1. Overcoming the challenge

Remember that preparatory work is important. Make contact with other systems that have already linked their data and gather case studies and use cases to inform conversations with organisation and system leaders. Consider holding events to engage with staff at all levels and stakeholders, so that everyone understands the benefits to their own organisation and to the wider system and communities.

Speak to operational managers and find out how to interpret data from the people who are using it. This will allow identification of any gaps in the data or issues where different coding or fields are used. Do not make assumptions about how it works and what it says.

Engage with your population and keep them and the organisations involved up to date with progress and successes. Put in the work to create good working relationships and make sure individual data teams are brought together via a steering or similar group to keep the project on track and to develop ideas for future questions to be answered using data linkage.

6.3. Resource

Finance is a challenge; however, data linkage needs both financial resource and expertise. It is important to demonstrate improvement through data linkage in order to secure the necessary funding.

The broader the data, the more expertise is needed to understand it at speed; while data linkage may be possible, data teams need to understand the data to avoid inaccurate analysis or interpretation. Linkage will always rely on humans looking at data to check that what the computer has linked is right, or to fix spelling or other data entry errors.

People who do use data linkage tend to see its benefits and want more from it, which raises another challenge of how to provide an equitable service to everybody and how to manage and prioritise change/development requests.

It is important that organisational buy-in at a senior level is translated into action at a grass roots level, with resources and protected time allocated within IT and data teams and the operational teams responsible for collecting and recording data. It is important to understand the other projects or organisational priorities which may be competing for the same resources, in order to ensure realistic expectations around the speed at which data linkage can occur.

6.3.1. Overcoming the challenge

In some areas, local integrated systems may be able to create single operating models to facilitate data integration and access to the data for all appropriate requests.

A cross-system intelligence function with multiple analytical teams collaborating, all using the same linked dataset, would be a real benefit to all involved.

The goal is to create a learning health system in which different organisations (or different analytical teams on behalf of their organisations) collaborate with each other in real time, combining their respective areas of expertise or specialisms to achieve the best outcome overall as a system.

6.4. Accessing and managing the data

There are a number of potential issues related to accessing and managing data. For example, levels of data sophistication and definitions tend to vary between different organisations.

Organisations working together need to consider accountability for the data and the consequences and implications if data linkage goes wrong.

There are computational challenges with working with bigger datasets, in addition to questions around how to link any dashboard to the existing clinical system used locally and navigating the governance process associated with that.

Where organisations want to integrate NHS Digital data, where NHS Digital is a data controller, there are rules to follow, such as the appropriate pseudonymisation process, the lawful basis for data linkage, the purposes for which the data may be used and the potential recipients of that data, and any limitations applicable to such data usage and sharing.

> *"People and organisations involved need to have confidence in the data – there must be trust that the data will be used responsibly and trust in the data itself."*

It is also important to remember that linking data from outside the NHS is not part of NHS Digital's agenda, so accessing data from the independent sector/private healthcare providers, while important, may be more complicated. It could also bring significant risks associated with commercial sensitivity when dealing with organisations which could compete with one another commercially.

Ongoing issues include the use of the NHS number in non-NHS sectors (local government and third sector), the public appetite for this, and public awareness of the benefits, which can be assessed through local PPI work.

Other considerations include deciding when person-level linkage is needed, when it is worthwhile to explore postcodes and how to resolve conflicts on which dataset takes primacy (e.g. where a person has given different genders/ethnicities in different datasets). These issues can be assessed via DPIAs or via a robust governance process for considering and approving requests for linkage or access to linked data which put considerations of necessity, proportionality, data minimisation and ethics front and centre. The right data should be made available to the right people at the right time.

More data add more complexity to interpretation, and more individualised reporting makes user engagement challenging.

6.4.1. Overcoming the challenge

It is important to ask the right questions at the outset to get organisational agreement to share data at a system level, rather than just at an organisational level, for example, for planning or research purposes. It's also essential to be clear on the use cases/research hypotheses.

It is better to have one lead team with responsibility for storing, analysing and managing the data, and bringing other stakeholders in. Identifying multiple different parties for this aspect of the work would become very convoluted to manage practically.

6.5. People and relationships

Managing relationships between the different organisations involved and getting appropriate buy-in can be a challenge. Data linkage only works when everyone buys in.

While some organisations have experienced a lack of board-level understanding about informatics and data analytics, it can also be time-consuming and challenging to get clinicians involved and manage issues related to organisational agreement.

> *"It's only when the value of data analysis is really understood that people involved with logging and processing the data will adapt the way they log and manage the data in order to ensure its quality."*

Building trust between providers (including local authority) and commissioners and exploring the contractual implications of sharing data are important. For example, if providers share data about capacity, they need to be confident it will not then be used by commissioners when renegotiating contracts or by competitors.

6.5.1. Overcoming the challenge

Building relationships is key. Through good relationships, it is possible to encourage people to make brave decisions, such as persuading chief executives of the relevant organisations to link police data to health data to try to target families at risk of domestic abuse. An ethics committee can be a central point to address any relevant concerns.

A cross-partnership team should lead and define questions; data has to be open to all partners to use. A steering group with data professionals and clinical representatives from all the involved organisations can work well. These groups should also include the public; for an example, see OneLondon.

If linking GP data, get primary care involved early on and ensure the local medical committees are on board.

A shared understanding

Co-production and co-design are important. Staff on the front line may use data capture tools and electronic health records in practical but idiosyncratic ways that do not make sense to analysts. It is vital to begin the conversation with everyone to determine what question is being asked and what the data means.

To support engagement, make sure you use recognisable language.

A shared understanding of the benefits of data linkage is also needed and can be achieved through signing up to a digital charter.

Ongoing support

Be aware that many people working on data linkage projects may well have been through several iterations of similar work in the past and could suffer from ‘project fatigue’. Make sure that once the dataset has been set up, you are able to provide ongoing support for and engagement with the users of the dataset to ensure the success of the project right through to completion.

7. Data linkage in action

The step-by-step guides in this section demonstrate examples of real-world application of data linkage.

7.1. Midlands and Lancashire Commissioning Support Unit

The business intelligence team at Midlands and Lancashire CSU has been working with Worcestershire County Council (WCC) on the development of machine learning to support the delivery of adult social care. A DSA to link health and adult social care data has been put in place between the two organisations plus external partner Predict X.

An initial grant was received from NHS Digital to support the linking of data after WCC put in a bid, with the support of the CSU to manage and bring in health data. The planned use of predictive analytics and AI meant also involving specialists at Predict X, whose aspirations fit well in terms of where the WCC and CSU wanted to go.

Work is currently focused on the Covid-19 response and recovery.

7.1.1. About the dataset

The dataset links current health and social care information and offers the ability to also flow in data from assistive technology partners in the future. It also links with a broader range of data, including some segmented datasets created using ACORN and other datasets such as the Covid-19 shielding patient list, the index of multiple deprivation, and mental health and community health services.

Data has been taken in from the last two and a half years for health and social care, which was then linked together using a range of pseudonymisation techniques. It was then passed on to Predict X, who applied AI algorithms to the data for a range of outputs before supplying those back to the WCC in a pseudonymised format.

The CSU business informatics team uses the Nottingham University Open Pseudonymiser to generate a unique SALT (security) key which is shared with WCC. WCC produces data in a clear format, then runs it through the Open Pseudonymiser tool applying the same SALT key; this then pseudonymises the identifiable fields. The CSU can then apply the same SALT key to its own datasets for secure data linkage; see Figure 3.

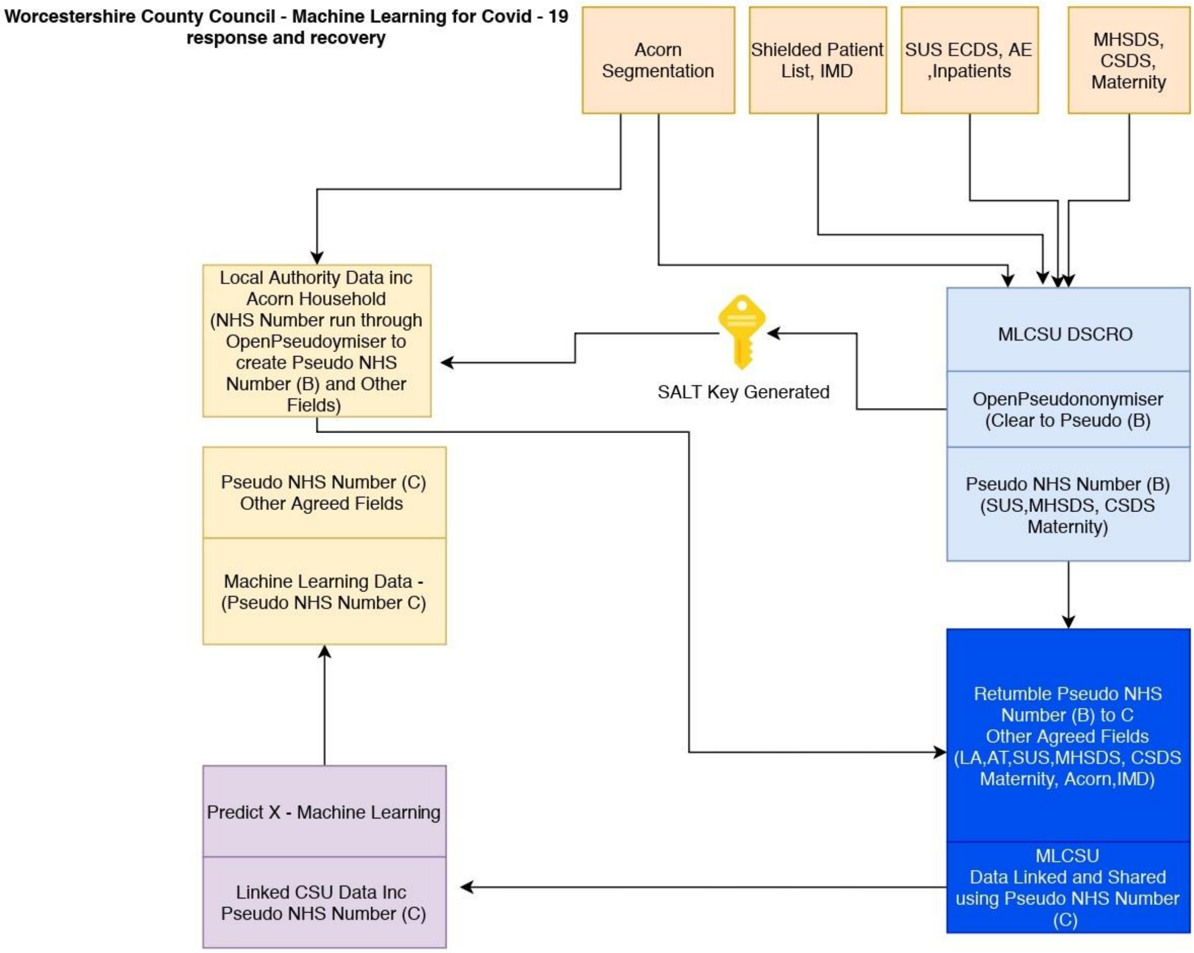

Worcestershire County Council - Machine Learning for Covid - 19 response and recovery
Acorn Segmentation
Shielded Patient List, IMD
SUS ECDS, AE ,Inpatients
MHSDS, CSDS, Maternity
Local Authority Data inc Acorn Household (NHS Number run through OpenPseudoymiser to create Pseudo NHS Number (B) and Other Fields)
SALT Key Generated
MLCSU DSCRO
OpenPseudononymiser (Clear to Pseudo (B)
Pseudo NHS Number (B) (SUS,MHSDS, CSDS Maternity)
Pseudo NHS Number (C) Other Agreed Fields
Machine Learning Data - (Pseudo NHS Number C)
Retumble Pseudo NHS Number (B) to C Other Agreed Fields (LA,AT,SUS,MHSDS, CSDS Maternity, Acorn,IMD)
Predict X - Machine Learning
Linked CSU Data Inc Pseudo NHS Number (C)
MLCSU Data Linked and Shared using Pseudo NHS Number (C)

Figure 3: Flowchart for WCC machine learning for Covid-19 response and recovery

As part of this project, WCC pulled out data focusing on domiciliary care, pseudonymised it and shared the resulting dataset with the CSU through a secure data transfer. The CSU has ‘file watchers’ to look out for any new versions of data landing in the CSU environment, who first understand what the dataset is and whether it matches the agreed specification, then pseudonymise the relevant fields such as NHS number, date of birth and postcode. They then run another pseudonymisation process in a consistent way in which the Secondary Uses Service (SUS) data, inpatients and A&E data has been processed, again facilitating reliable linkage.

7.1.2. Additional layers of data security

Once linkage has been achieved, the CSU completes another pseudonymisation exercise before the data is sent to another third party (Predict X) to reduce any risk of reidentification by re-tumbling the data through the open pseudonymisation tool. This means that when the data is returned to WCC, it is not possible for the data to be reidentified.

The CSU chose this method for data sharing because of the NHS Digital drive to minimise any risks in using patient-identifiable data. Pseudonymisation at source is preferred before data is transferred to any other organisation; the method has been tried and tested and is simple to implement.

The linked dataset has been up and running for 12 months and is refreshed on a monthly cycle, coinciding with the monthly SUS and local authority updates. It does not currently include primary care data, but this could be a future consideration if it is found to be beneficial and appropriate governance is in place.

7.1.3. Key learning and advice for others

- Start work early on the DPIA and IG to help you identify the planned use of the data and the purpose for the linkage; understand the security layers you need to put in place.
- Ensure partners understand the processes so that if you want to expand work in the future, they already have confidence in the system.
- Agree the key fields you need to include so that your outputs match what you are setting out to do.
- Only request the necessary data. The more data you link, the higher the risks, and you need to be able to explain the purpose of the data you use. For example, if you don't need to know the patient's location, don't include it as a field (or just use a partial reference), and try to remove any free text that may include personal details.
- Give stakeholders assurance around the legal basis for the linkage.
- Funding may be an issue; demonstrate the real value to partners, to encourage buy-in.
- Understand your key purpose. For example, WCC wanted to reduce the number of patients attending A&E by putting social care packages in place, and needed to know what they had to put in place to address that purpose. The CSU knew what datasets existed, but if the analysis showed a gap or a key measure that had not been included, they could revisit the agreement to include other data sources.

7.1.4. Next steps

Key for WCC now is access to linked data where they can easily expand the use cases, understanding the methods and what needs to be done.

The ICS is key to the future development. Data linkage is not an end in itself. Its benefit comes from the collaboration it creates; the data is just a lever. The aim is to turn linked data into actionable insight using AI.

The data linkage team is now working with partners such as social care to improve their internal processes, so that data is consistently processed and can go through the automated processes.

For more information, contact Phil Rowley, Head of Business Intelligence – Data Management, Midlands and Lancashire CSU, at phil.rowley@nhs.net.

7.2. Kent Integrated Dataset

The Kent Integrated Dataset (KID) includes health and care data for more than two million people from 240 GP practices, acute trusts, adult social care, mental health services, public health and community health, plus a range of other organisations. It was started in 2016 to support the Year of Care programme by defining an overall cost of the care people receive for long-term conditions, and developed from there.

The KID was built by the Data Warehousing Shared Service at Maidstone and Tunbridge Wells NHS Trust (HISbi), which already held data from a range of NHS trusts and other organisations. The KID team quickly realised that it had much wider potential uses.

The KID was frozen in 2019 but remains a valuable resource for researchers. The team is now developing the Kent Research Network for Education and Learning (KeRNEL) linked data facility, using lessons learned from the KID.

7.2.1. About the linked dataset

When it was first built in 2016, the dataset took in the databases already housed in the data warehouse which had the correct governance in place. Data was then gradually added from other sources as and when agreements were put in place. The decision to start with the more robust data gave stakeholders confidence in the dataset and allowed the system to get up and running as quickly as possible.

The main method of data linkage was through NHS number, with health and care data flowing into the linked dataset having been pseudonymised at source.

A black box algorithm written by the team gave individuals unique IDs, which meant that the data could not be de-anonymised and that organisations submitting data could not match their records with the linked data to identify any individuals.

Where data was not recorded at the personal level (for example Fire and Rescue data), it was linked at pseudonymised postcode/UPRN level. The Experian MOSAIC tool was used to assign pseudonymised UPRNs to the data, which could then be linked with record-level data from other organisations, as UPRNs could be assigned to all data.

Organisations not involved in the data warehouse were provided with an API (or ‘widget’) to allow them to submit data to the KID that was automatically pseudonymised.

7.2.2. Governance

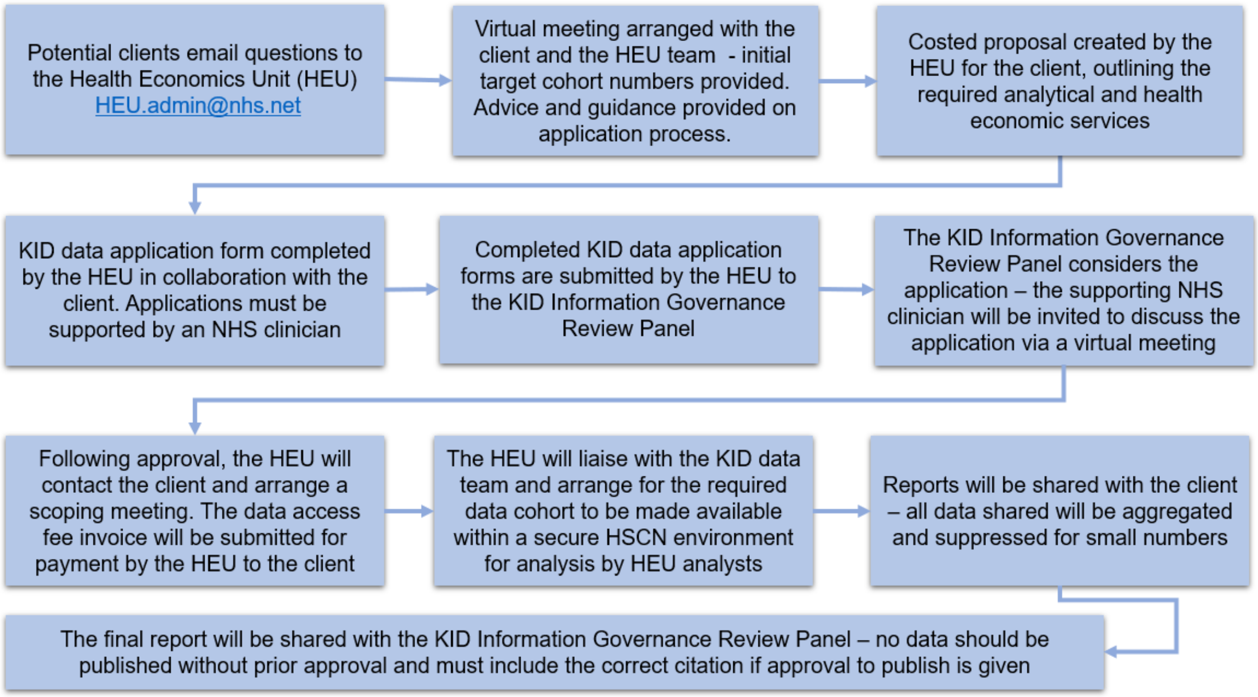
No national guidance on data linkage existed when the KID was created, so the team worked closely with senior NHS figures during development of the governance framework.

Every organisation taking part signed a memorandum of understanding around the KID in general plus a contract and data processing agreement covering the governance.

GPs needed extra agreements to allow data to move out of their practice, which they signed with the supplier who pushed the data out.

7.2.3. Access to the linked dataset

Initially, only county council staff or those seconded to the council were allowed direct access to the linked dataset. However, this was found to be very restrictive. The Kent and Medway CCG took ownership of the KID in March 2021, and wider access will be allowed going forward, with a rigorous approval process overseen by the Shared Healthcare Analytics Board – a group comprising all organisations across Kent & Medway that reports into the CCG (Figure 4).



Potential clients email questions to the Health Economics Unit (HEU) HEU.admin@nhs.net

Virtual meeting arranged with the client and the HEU team - initial target cohort numbers provided. Advice and guidance provided on application process.

Costed proposal created by the HEU for the client, outlining the required analytical and health economic services

KID data application form completed by the HEU in collaboration with the client. Applications must be supported by an NHS clinician

Completed KID data application forms are submitted by the HEU to the KID Information Governance Review Panel

The KID Information Governance Review Panel considers the application – the supporting NHS clinician will be invited to discuss the application via a virtual meeting

Following approval, the HEU will contact the client and arrange a scoping meeting. The data access fee invoice will be submitted for payment by the HEU to the client

The HEU will liaise with the KID data team and arrange for the required data cohort to be made available within a secure HSCN environment for analysis by HEU analysts

Reports will be shared with the client – all data shared will be aggregated and suppressed for small numbers

The final report will be shared with the KID Information Governance Review Panel – no data should be published without prior approval and must include the correct citation if approval to publish is given

Figure 4: The current process for researchers such as the HEU to access the KID

7.2.4. How the KID inspired the KeRNEL

The KeRNEL will expand the breadth and depth of the KID, representing a journey towards incorporating more wider determinants of health and evolving data linkage methodology.

The team is looking at pulling in more in-depth health data, such as early warning scores for clinical deterioration, as well as reaching out more widely to other organisations; for example, to source data on free school meals and air quality.

7.2.5. Changes made based on the lessons from the KID

Unlike the KID, the KeRNEL receives data into a central location, with relevant data being pushed out into databases specific to individual projects as needed.

The data flow has been redesigned to use a pseudonymising method that will allow both the de-personalising and reidentifying of individuals, which is key for buy-in from researchers and GPs. Reidentification means that if a piece of research identifies a person as being particularly at risk of a negative health outcome, their GP or other approved clinician could alert them to this risk and offer an appropriate intervention. There is now less resistance to sharing for this reason, but more care is needed on the IG side to instil confidence.

Plans are in place for the KeRNEL to be updated at least every night so that it can be used for performance and planning as well as for research, such as evaluating an intervention (researchers will be able to separate out a patient cohort that has received a particular intervention and compare it with another that has not).

7.2.6. Advice based on the KID experience

- Funding is key; data linkage should not be attempted on a shoestring.
- Staff should be allocated to the data linkage programme, not asked to incorporate it into a wider day job.
- Buy-in from the health community is important. A linked dataset needs to work for the wider community so you do not meet with objections.
- All the governance and other building blocks should be in place before you bring the data in.

- Make sure you have your purposes agreed, with high-level sign-off, and agree who will be the owner/lead (e.g. CCG/council).

For more information about the KID or KeRNEL, email Peter Gough, HISbi Head of Service, Maidstone & Tunbridge Wells NHS Trust, at peter.gough@nhs.net or go to https://kmkernel.org/.

7.3. Care City

Care City, a community interest company, was founded in 2019. As part of its development, partners including UCL Partners, Care City, North Thames CLAHRC, BHR CCG,[14] the London Borough of Barking and Dagenham and North East London Foundation Trust (NELFT) created a unique linked dataset: the Care City Cohort.

A key priority for the linked data was to better understand population health, exploring the themes from Marmot and others on social context and long-term health needs.

7.3.1. About the linked dataset

The dataset contains individual and household-level linked data across the partner health services and Barking and Dagenham (B&D) council, including:

- Sociodemographic and health (e.g. age, gender, ethnicity, smoking status, BMI, long-term conditions)
- Where individuals live (e.g. levels of deprivation, household occupancy, household tenure)
- Health and social care service use (e.g. A&E attendances, GP contacts, social care packages, mental health inpatient stays)

The dataset is hosted in the Barking and Dagenham, Havering and Redbridge (BHR) CCG Data Safe Haven, with different datasets linked together using linkage keys in place of NHS numbers. Pre-existing linkage of primary and hospital data housed by BHR CCG serves as the starting point for the wider linkage (see Figure 5), with data from other sources linked later.

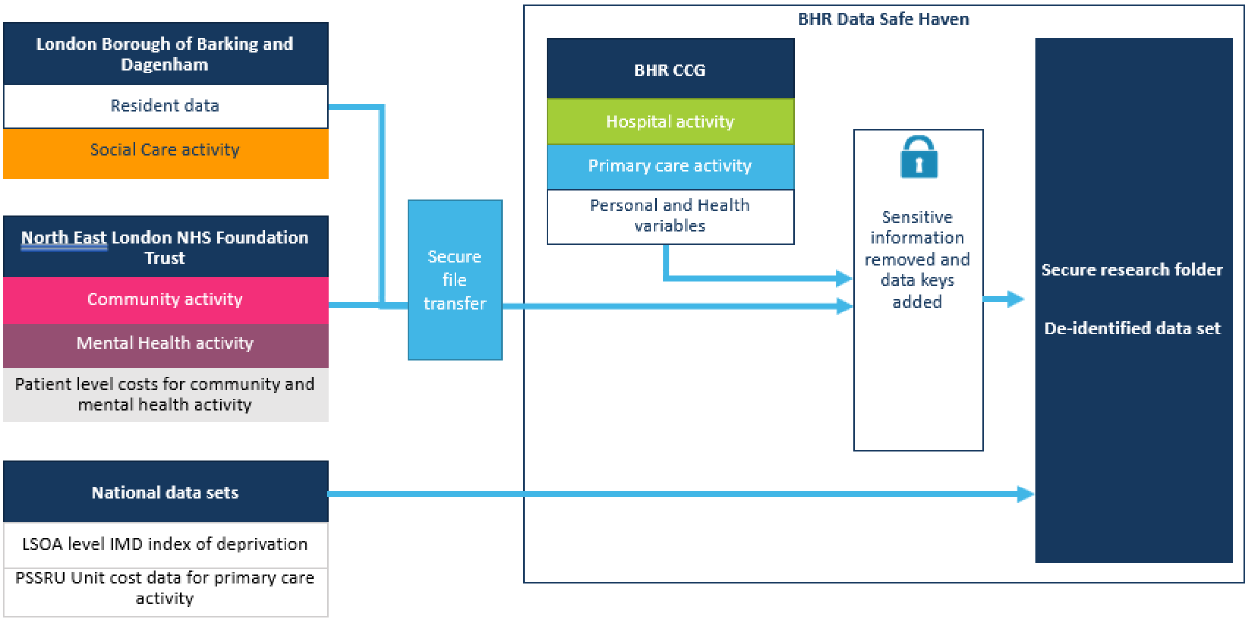

London Borough of Barking and Dagenham
Resident data
Social Care activity
North East London NHS Foundation Trust
Community activity
Mental Health activity
Patient level costs for community and mental health activity
National data sets
LSOA level IMD index of deprivation
PSSRU Unit cost data for primary care activity
Secure file transfer
BHR Data Safe Haven
BHR CCG
Hospital activity
Primary care activity
Personal and Health variables
Sensitive information removed and data keys added
Secure research folder
De-identified data set

Figure 5: Data flows for activity and resident information for residents of Barking and Dagenham from 1 April 2011 to 31 March 2020 who were registered with a Barking and Dagenham or Havering GP practice

[14] On 1 April 2021, NHS Barking and Dagenham CCG was merged with NHS City and Hackney CCG, NHS Havering CCG, NHS Newham CCG, NHS Redbridge CCG, NHS Tower Hamlets CCG and NHS Waltham Forest CCG to form NHS North East London CCG.

7.3.2. Dataset design and data cleaning

SQL Server is used to pull the dataset together, with extract files created in Excel. All data is de-personalised, and a unique property identifier (in place of the UPRN) is used to facilitate household-level analysis.

Resident and social care data about the social context of each individual are taken as a yearly snapshot on 1 April, which is considered sufficient for research; all hospital and service use activity is used to give a view over time.

The dataset includes residents of the borough plus others who are registered with a B&D or Havering and Redbridge GP, as the data was already available. It does not include people not registered with a GP, which does present limitations.

Data cleaning is essential. The team found a number of issues; for example, 250 children with a special educational need (SEN) appeared twice in education records (once with the school that initiated the SEN process and once with the receiving school). Duplicates are now removed each year. Data gaps in social care were also found; for example, inconsistency in recording start and end dates of social care packages. Cleaning the data exposed design decisions needed on the dataset and how it works.

The data cleaning rules used by the council have been updated and refined over time. These include using probabilistic logic matching in cases where, for example, a record does not include an NHS number.

7.3.3. Access to the linked dataset

B&D council and the CCG each have embedded researchers with direct access to the linked dataset; analysts within UCL can also access the dataset.

Others including researchers, teams from the local health and care system and teams from other geographies that want to collaborate with or learn from Care City can apply to these teams to carry out an analysis for them or to request direct access. The team, along with academic reviewers, assess the suggested protocol to ensure suitable methods of research are proposed and to identify how the output could be of use to B&D.

Researchers must use R Studio for analysis, as it is open source and allows others to replicate projects on their own datasets.

7.3.4. Information governance

Bespoke IG agreements were set up between NELFT and BHR CCG for the data flow of mental health and community services, and with B&D council and BHR CCG for council and social care data flows. The full linked dataset proposal was approved by the IG committee, which included all the GPs from the borough. The agreements cover a two-part process of getting the data in and approving its use as part of a linked dataset.

7.3.5. Engagement and involvement

The Care City Community board has a critical role in providing public perspectives on the use of the dataset and supporting the partners to ensure findings are relevant and accessible to all.

7.3.6. How the linked dataset is being used

Care City is taking part in the Health Foundation’s Evidence into Practice programme. The aim is to use the linked dataset to better understand how the population uses services alongside wider social determinants of health, and to translate research findings into actionable insights for service staff, system leaders and local policy makers.

7.3.7. Projects so far:

- Access to vaccines in Barking & Dagenham
- Do care homes have one GP or many?

- Domiciliary care and hospital discharges

Other projects have looked at:

- Using linked health and resident data to see any clustering of people likely to need to shield during the Covid-19 pandemic, so that the council could target support geographically
- Understanding service use patterns in people in the last year of life
- Understanding service use patterns in different care settings
- Whether people with a carer have different levels of service use across different settings compared with those who do not have a carer but have similar characteristics

For more information, email Jenny Shand at jenny.shand@uclpartners.com or go to www.carecity.london.

7.4. Connected Health Cities

Cheshire and Merseyside's PHM platform, CIPHA, was established in the space of three months to support a coordinated health and care system response to the coronavirus crisis.

It was clear that the region needed a PHM platform to link data and make it available for public health colleagues. A review of existing data sharing and the technologies used across the region took place, including discussions with other areas such as Manchester.

The CIPHA team then worked with software provider Graphnet (a shared care record system already deployed across England, including in Manchester) to build a new core platform and linked dataset covering GPs, hospitals and some social care data. Tiered DSAs and a governance framework were already in place as a part of existing data sharing activity. Combined with the control of patient information (COPI) notice, this enabled the team to accelerate the work to integrate the data but still required meeting with individual data controllers.

Dashboards were then created that could help to inform local policy and decision-making at gold, silver and bronze levels of the local pandemic response command structure.

7.4.1. Use cases for CIPHA

A smart testing programme was also launched in November/December 2020 for lateral flow testing, which included data from the PHE Second Generation Surveillance System feed and the demographic, risk factor, treatment and outcome information for patients admitted to hospital with a confirmed Covid-19 diagnosis, as recorded in the PHE Covid-19 Hospitalisations in England Surveillance System.

The team were also permitted to add in vaccination data, allowing them to get more timely updates that were helpful in targeting vaccine-hesitant populations. They also took part in the events research programme introduced by the Department of Culture, Media and Sport, tracking cases linked to large-scale events such as nightclub evenings and a special event at Sefton Park. Through the data linkage and with attendees' permission, they were able to identify where people were testing positive in order to manage potential outbreaks in real time, which helped to inform local policies.

Current use cases for CIPHA, deployed through the current expansion with other regions, include pulse oximetry to manage Covid-19 patients virtually, a Covid radar to better understand the impact of Covid-19 in communities, BP@home (management at home for people with high blood pressure) and discharge management. Projects are also in development for waiting list management, early intervention for advanced social care support and telehealth.

7.4.2. About the dataset

The data is livestreamed where possible and comes in from GP practices each night. There are some differences in local system availability or capability that need to be taken into account; for example, one organisation uses CSV files for some of its data and sends them by secure FTP, and is being supported to move to a more robust data transfer system (extraction, transformation and loading (ETL) for SQL Server).

Data matching is mostly carried out through the NHS number, with probabilistic matching used where this is not available. For example, probabilistic matching was needed for the events research programme because data from venues did not include NHS numbers; a combination of name and address was used, taking into account misspellings, and much of that work was completed manually. Using a combination of data from the testing service and demographic data from the NHS Spine – probabilistic matching – the team was confident of its rate of around 95% matching.

7.4.3. Expanding the CIPHA team

The CIPHA team, hosted by Mersey Care NHS Foundation Trust, includes staff from a range of organisations across the area, including university colleagues with honorary NHS contracts that allow them to prepare data. The Graphnet team has contributed to the development of an OpenSAFELY open-source software stack which aims to expand access to the linked data and analyses.

There are a number of working groups dedicated to the expansion of CIPHA, rolling out the programme to a population of more than 16 million citizens. It currently covers 12 system areas, including Lancashire and South Cumbria, Surrey Heartlands, Northamptonshire, and Manchester Community Central health and care partnerships.

7.4.4. Next steps underway

DSAs are now in place to support use of the linked data in direct care, provided through Graphnet, and population health agreements are being updated with the GP population to create a trustworthy research environment to sit on top of the dataset. The aim is to create a solution that is more robust and sustainable, supporting the PHM research agenda and the bigger questions for analysis.

Using extension funding from NHSX, the CIPHA team will also work with other areas using Graphnet, looking at their use cases to see what is transferable and networking across regions.

7.4.5. Addressing the challenges

There are a number of learning points from addressing the challenges in Cheshire and Merseyside:

- Variance in coding between different organisations remains a challenge, although this has been helped in some areas by the need for quality outcomes framework reporting. A standardised data dictionary between sites is useful.
- Linking data exposes gaps in analyses. This is where locality/regionality of data and analysis is important. At a national level, it is difficult to get an idea of differences and variables, but at a local level it is possible to fix issues and ask why variables have suddenly changed.
- Technology is not the problem for linking data; a clear purpose, well-defined IG and time to ensure all stakeholders are on board are needed.
- The pandemic served as a driver for CIPHA, but the people involved were already aware of governance requirements and had agreements in place, with documents in a form they had seen before. Covid-19 also acts as a single use case with clear objectives.
- As we move to restoration of services, we must continue to give back to the people who are providing the data, meet their requirements, and make dashboards and tools available to them.
- We must also make it clear how the data around PHM can help GPs in one-to-one care and that it is not about performance management or finding efficiency savings.

For more information, email gary.leeming@liverpool.ac.uk or go to www.cipha.nhs.uk.

7.5. Surrey Heartlands ICS

The Surrey Heartlands ICS is currently developing a comprehensive data strategy which recognises the need for data sharing and integration across partners. The work is divided into four quadrants which each look at the data through different lenses on why and how data is brought together, how and where it is aggregated and who has access to that data. Progress is at different stages of development in each quadrant.

7.5.1. The quadrants

- Sharing patient information between professionals to successfully deliver direct care and enable collaboration around the needs of the citizen.
- Secondary uses for direct care data – understanding the impact of specific interventions on health outcomes and planning services from GP practice up to integrated care provider level for frailty services etc. As a wave 1 population health management ICS, this also feeds into the PHM strategy with a possible extension into the full remit of research.
- Quality data – which kinds of data colleagues in quality and multidisciplinary professionals need in order to understand long-term outcomes. How we can correlate outcomes and get data into an environment where we can pose hypotheses as well as testing a cohort of patients.
- Financial data as a system and as a commissioner.

7.5.2. Linking data for direct care

The Surrey Care Record has been introduced using the Graphnet platform, which links all the ICS partners, including community and mental health providers, local authorities, GPs and four acute trusts. This has been delivered on the basis of explicit DSAs, without needing to call on the powers within the recent control of patient information (COPI) notice, ensuring that all the processes are as transparent as possible.

The care record has been linked to the Thames Valley and Surrey (TVS) Local Health Record, which is crucial as 10% of Surrey patients flow across into the neighbouring Frimley ICS area which the TVS record also covers, and the local authority is co-terminus with Surrey and Frimley. To get an overall picture it is important to see both areas, and therefore DSAs are in place with Frimley.

Graphnet pulls the data in, which is then verified and deduplicated, cleaned and made available to analysts and clinicians.

7.5.3. Secondary uses for direct care data

Through links with the TVS health record, the Surrey Heartlands ICS has access to a population health analytics platform that enables it to analyse local data rather than using national and SUS datasets that are up to eight weeks out of date. The TVS population health management dataset and its regional footprint unhook all the rich PHM capabilities that come with the platform, and this is very important to the ICS for risk stratification and predictive analytics.

As part of the developing strategy the ICS is looking to bring in data from the police, voluntary sector, districts and boroughs for housing and air quality, for example, to provide richer data for PHM.

7.5.4. Software solutions

The ICS uses a range of software solutions, including Alamac and Beautiful Information. These platforms are used to take regular snapshots of activity data, combine that with SUS data and allow analysts to look backwards and establish causation – which is helpful, but not forward looking.

For that reason, the team aims to use the linked dataset for waiting list management, looking at how they can combine waiting lists around the local trusts, flow cases around the system, consolidate and get economies of scale, and segment the waiting list to focus on a specific population. At the moment such work is very manual, but the team wants to create a command centre to see what is happening on their patch and make quick decisions.

They have outsourced core data processing to North East London CSU, who also manage core data processing with acute providers. The aim is to get data into an environment where analysts can look for the richness in the data to ask new questions; the ICS often answers questions set by NHS England, but they also want to identify and ask their own questions.

7.5.5. Key learning and advice for others

- Start with the strategy. Try to understand core questions your data needs to deliver.
- Get a really good sense of which organisation the data sits with.
- Agree data principles of how the linkage will work and get buy-in before moving towards the detail of your data environment.
- Keep your feet on the ground in terms of ambitions.
- Get your IG and sharing agreements right. The ICS is passionate about being explicit about any uses of people’s data: “If we are going to build any kind of digital services, citizens need to trust that we are doing the right thing with their data.”

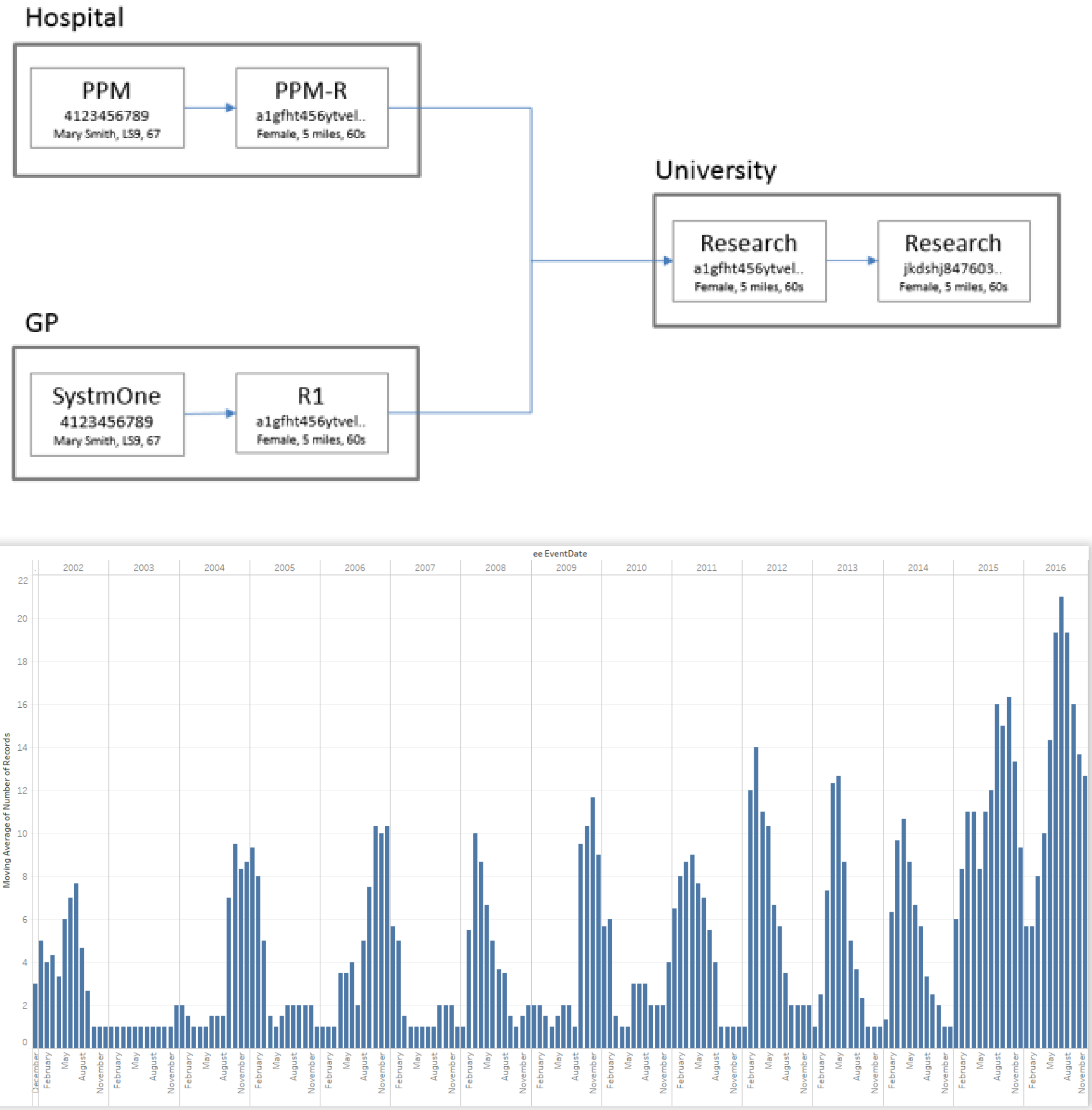
For more information, contact Katherine Church, Chief Digital Officer, Surrey Heartlands ICS, at katherine.church@surreycc.gov.uk.

7.6. Health Data Research UK and DATA-CAN: comprehensive patient records, Macmillan and Leeds

Since 2018 a team from Health Data Research UK (HDRUK/DATA-CAN) has been working on a groundbreaking programme in Leeds which brings together secondary and primary care data to analyse patient journeys for people receiving an urgent cancer referral.

Known as the comprehensive data patient care records for cancer outcomes,[15] the resulting dataset looks back at 10 years of the medical history of cancer patients before their diagnosis and treatment, their long-term outcomes, and the medical history of matched individuals without cancer who form a comparator cohort.

The aim is to find patterns of relapses and see whether warning signs can be picked up earlier, raising the chances of earlier diagnoses and better patient outcomes.



Cancer recurrence

Figure 6: The data linkage and resulting analysis

[15] https://web.www.healthdatagateway.org/dataset/ce13db83-cedb-4ff9-9f6d-fe668d872da4

7.6.1. About the project

Much data analysis in the NHS relies on episodic measurement in secondary care without any contextual data. The HDRUK team considered linkage with primary care data as essential to creating a wider health context and a longitudinal view. They formed a collaboration with TPP/SystmOne (a Leeds-based primary care software provider used by most GP practices in the area) and worked with them to create a data repository using OpenSAFELY software.

The data repository, known as ResearchOne (R1), houses an extract from across all Leeds patients, linked together using Open Pseudonymiser. The data is derived from linked primary, secondary and tertiary care electronic health records and participant survey responses. Data is de-personalised at source (Leeds Teaching Hospitals NHS Trust (LTHT) and R1) and linked using matching pseudonymous digests that are re-pseudonymised upon linkage by University of Leeds IT to produce irreversible pseudonymous data that is processed into a research dataset.

The data relates to 431,352 patients in the UK with whom LTHT has a 'legitimate patient relationship' and who were determined by LTHT to have had a cancer diagnosis between 2004 and 2018 or to be a matched non-cancer control.

Subsequent data refreshes (currently annual, but possibly moving to quarterly now the system is automated) go directly to the Yorkshire and Humber Cloud on the Google platform, which is now used to house the repository, and the team has placed the metadata specification onto the HDRUK gateway so that other researchers can see how it may be able to help their research.

7.6.2. Ensuring high data quality

The team carried out an in-depth data quality assessment to check that the integrity of the data continued throughout the linkage – making sure that the pieces of data being added and linked did not change the story that was being told. A quality assessment repeated on the 30 or more tables of data proved it to be of a high enough standard but also enabled the team to flag up any issues with feeder systems about any gaps in the data and to make suggestions for writing new scripts.

Through this continuous quality development and improvement, HDRUK developed a data utility framework that measures the metadata to say how useful it is to researchers and for further data analytics. This work was published in the BMJ Health & Care Informatics Journal in May 2021 and the framework has been adopted across HDRUK. The team have been able to demonstrate a real benefit not just from a research point of view but also in informing pathways for direct care.

To find out more about the HDRUK work, click here.

7.6.3. Uses of the linked dataset

Geoff Hall, Professor of Digital Health and Cancer Medicine at the University of Leeds and Chief Clinical Officer for Research, LTHT, who specialises in gynaecological cancers, has been able to use the dataset analysis to see patterns of relapse with his patients, which could enable clinicians to pick up patterns and symptoms earlier. His team has been able to start some predictive analysis around pathways for people with relevant cancers, through the unlimited processing power of the Google Cloud platform.

Professor Hall's team has access to 1.8 million records and can look at the characterisation of the cohort and see any patterns. The in-depth quality assessment allows them to have confidence in the data they are using.

By looking at the patterns being revealed throughout primary, secondary and tertiary care via data linkage, clinicians will hopefully be able to intervene much earlier by looking at presenting conditions that would not ordinarily be considered for such cancers; for example, symptoms and behaviours related to digestive issues, such as people taking antacids. Clinicians can also look at survival curves and the impact of comorbidities such as diabetes (Figure 7). By using the data, they can see the effects of comorbidities and adapt clinical practice to provide better clinical interventions at an earlier stage, resulting in a better prognosis.

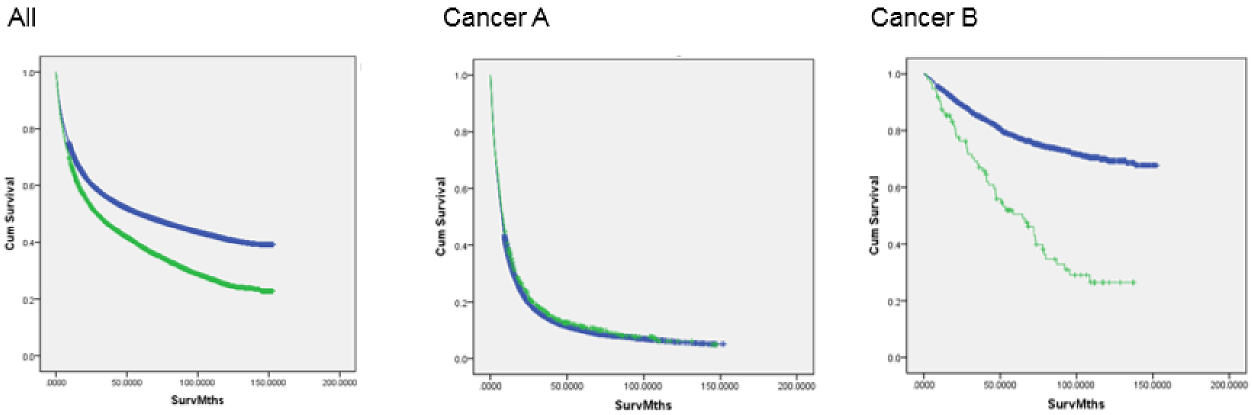


Figure 7: Analysing the impact of diabetes on cancer survival

Having the data available through the Yorkshire and Humber Cloud allows clinicians immediate access to visualisations in Google Tools and access to data using Tableau.

Click here to find out more.

7.6.4. Next steps

Through DATA-CAN, the team is now working with NHS Digital on the national data from patients over a 10-year period to investigate the impact of Covid-19 on cancer. They believe that in Leeds alone there have been around 1,500 missing diagnoses, which scales up to around 85,000 missing patients nationally who are either not coming forward or are being picked up through later diagnoses.

By creating a longitudinal dataset, it is possible to see patients' journey across primary and secondary care rather than just an acute snapshot of their journey. The team hopes that through an extension of trusted research environments such as theirs, commissioners and providers will have access to data linkage that is already complete, removing organisational silos and allowing for better contextual data for pathway planning.

Standardisation of data quality and proving an infrastructure that allows for well-curated datasets to be held in a secure environment to be used by all, with all the relevant IG in place, will also be a great benefit.

For more information, email Monica Jones, Chief Data Officer, HDRUK Hub for Cancer DATA-CAN and Associate Director, HDRUK North Better Care Partnership at monica.jones@nhs.net.

Take-home messages

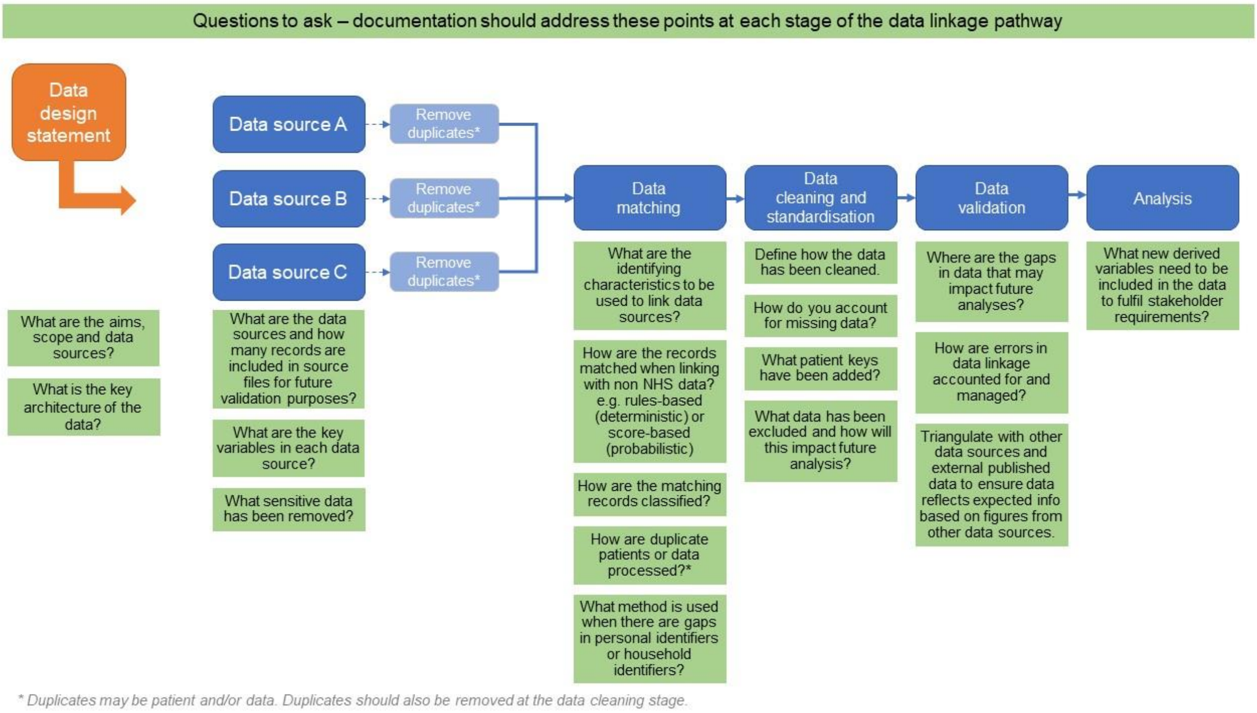
- Data linkage is a process of identifying, matching and merging records that correspond to the same person (or population) from several datasets. It offers a more granular level of detail about the actual health needs of a population in real time.
- There are many aspects to take into account before starting data linkage; it should not be rushed into.
- The quality, completeness and consistency of the data used are key.
- There may be times when using data linkage is not appropriate.
- Key enablers to data linkage include a clear understanding and support at a senior leadership level.
- Having a data sharing agreement in place with the data controllers and/or NHS Digital is essential. Establishing that agreement can be a lengthy process.
- A successful data linkage project involves managing not only the technical aspects but also relationships between numerous individuals and organisations.

Appendix 1: Technical guide

This technical guide covers the steps to successful data linkage. Some of the actions needed will vary depending on the IT/data analysis capability and capacity each local system has, but this guide provides a recommended pathway through the process.

The diagram below shows the key steps and questions you should ask yourself at each stage.

Data linkage high-level pathway



Questions to ask – documentation should address these points at each stage of the data linkage pathway

Data design statement

What are the aims, scope and data sources?

What is the key architecture of the data?

Data source A

Data source B

Data source C

Remove duplicates*

What are the data sources and how many records are included in source files for future validation purposes?

What are the key variables in each data source?

What sensitive data has been removed?

Data matching

What are the identifying characteristics to be used to link data sources?

How are the records matched when linking with non NHS data? e.g. rules-based (deterministic) or score-based (probabilistic)

How are the matching records classified?

How are duplicate patients or data processed?*

What method is used when there are gaps in personal identifiers or household identifiers?

Data cleaning and standardisation

Define how the data has been cleaned.

How do you account for missing data?

What patient keys have been added?

What data has been excluded and how will this impact future analysis?

Data validation

Where are the gaps in data that may impact future analyses?

How are errors in data linkage accounted for and managed?

Triangulate with other data sources and external published data to ensure data reflects expected info based on figures from other data sources.

Analysis

What new derived variables need to be included in the data to fulfil stakeholder requirements?

* Duplicates may be patient and/or data. Duplicates should also be removed at the data cleaning stage.

Before linking any data

Before beginning the technical work on data linkage, you need to consider a number of issues. These include agreeing the purpose and timescales for the project, the datasets needed, the data matching approach and the specifications.

Step 1: Get partners on board and partnership working in place

- Agree with all potential partners what you are trying to achieve and why. Prepare case studies about the benefits of linked data to help get partners on board.
- Decide what you want to measure and agree to link for that purpose and nothing further.
- Define which datasets you want to use.
- Engage with your in-house IT team early if, for example, database administration skills are required or significant setup and business-as-usual work will be expected, to make sure the data linkage ambitions are realistic.
- Ensure you have the data management and analysis resources needed to store, process and analyse this data. This includes a lead team with the relevant skills and expertise in cleaning, matching and linking data. This could be a local team already involved in data warehousing for organisations in the area or could include external suppliers of data management and IG (see Kent Integrated Dataset.

Step 2: Establish your information governance

- Once you have agreed the parameters of your linked data, establish an IG working group which includes clinicians from across the relevant areas and PPI representation.

- Form a joint data controller group and develop clear governance processes, including for situations where mistakes are made.
- Put a clear process in place for approving access to and use of the linked data, and address legal concerns on confidentiality from GPs and other partners.
- Use available tools to manage the IG process; for example, software to manage signing of an ISA, DSA or JDCA. Click here for an example JDCA and here for the generic NHSX template. Explore the use of data institutions to link and broker access to novel datasets or datasets where there are issues around trust or governance.
- Make sure the right contracts are in place between data processors and controllers and any other organisations involved. Where possible, try to use NHS standard terms with data processors, as the Provider Data Processing Agreement terms are very robust.
- Establish clear protocols on the approach that can be published and shared.
- Ensure all analysts have the required IG training and are aware of their obligations.

Step 3: Confirm your data/data management requirements and how they will be delivered

- Define data responsibilities, including who owns what. This may need to be included in your ISA, DSA, JDCA and/or contracts.
- Decide how often you will need to incorporate data – will you need live feeds or is new data needed less frequently? Your purpose for data linkage should answer this question (e.g. data used mainly for research could be ‘frozen’ in time, whereas data used for day-to-day system management will require frequent updates).
- Discuss with the integration team managing the data how the data management requirements will be achieved. A range of mechanisms could be used, predominantly FHIR APIs and SFTP transfers. Some IT systems can perform native data extracts/reporting, while others may require the use of a local integration engine to extract the data from source systems, process it and then share it on.
- Establish a system subscription for maintaining data quality and completeness by distributed/collaborative effort from the local organisations supplying data; the responsibility cannot be shouldered entirely by the data integration/warehouse team. GP practices may have difficulty resourcing data quality checks on their own depending on their size and capacity, but if they are working as primary care networks this should be possible. GPs can use templates within their systems to improve the completeness and accuracy of data capture to improve data quality. Templates for GPs using EMIS or TPP are available from providers such as Ardens. This issue should ideally be agreed as part of the planning phase, with areas such as data alteration/improvement, timing of data quality work and what to do with inaccurate data all agreed. There is no right way of doing this, and different projects will have different acceptance levels. Spine integration or batch NHS number tracing both assist with accurate demographics.

Linking the data

Step 4: Clean your source data/remove duplicates

- Start this process with datasets that follow a standard; use consistent files that have been confirmed across all the providers. If this is not possible, ask the local data teams to work together to deliver better data quality and consistency.
- Consider whether the data needs to be identifiable, pseudonymised or anonymised. You may need to pseudonymise the data at source.[16]
- Software solutions to consider:
    - Remove duplicates – use software such as Python.

[16] ‘Pseudonymising at source’ involves replacing personal identifiers with a pseudonym/reference number before data is shared. As the data cannot be identified, this approach removes the need for patient consent or other legal provision under the Data Protection Act or the General Data Protection Regulation.

  - Use NHS Digital tools for de-personalisation (removal of personal identifiers in the data). Consider whether you will also need the ability to de-pseudonymise the data to allow research results to be followed up in practice (see Kent Integrated Dataset).
- Create a data dictionary[17] that is meaningful across all sectors and is jargon free; do not assume that everyone will know what every data field contains. Click here for the NHS data dictionary as an example.

Step 5: Create the linked database – data matching

- Take in all the datasets, using FHIR, API, SFTP or other ETL processes, index them and create an MPI. Datasets are usually saved using SQL Server, either using cloud-based services or on in-house equipment. Software such as Python can be used for the process of linking data.
- Be clear on the type of matching you are using (deterministic or probabilistic 'fuzzy' matching) and consider checking manually, with input from relevant clinicians, whether the data looks 'right'.
- Be clear on what different datasets are being linked and make sure the data item being used for linkage is well populated.
- For health records, it is recommended to start first by matching through the NHS number if possible, and then use postcodes/addresses. It may also be possible and/or appropriate to bring in patient demographics.
- Make sure you have high-quality structured data; focus on data either captured using clinical terminology (e.g. SNOMED) within the patient record or subsequently coded into classifications such as ICD-10.
- Identify an MPI source early, as this could have a significant impact on IG and datasets. The MPI should refer back to where a patient is registered and the address, and should be the main dataset into which other datasets are integrated.
- The Experian MOSAIC tool can be used to assign pseudonymised UPRNs to data, which can then be linked with record-level data from other organisations.
- In the linked dataset, layer up demographic data that could change over time or could be different in different datasets using SQL or other relational database.

Step 6: Data cleaning and standardisation – accounting for and mitigating varying data quality and profiling

- Be clear on whether you are using deterministic or probabilistic data matching, or both.[18] It is important to be aware of the potential to draw conclusions about one person based on somebody else's behaviour or activities.
- Make sure time spans are the same when comparing datasets (e.g. do not link calendar years with financial years).
- The order in which datasets are linked is important. Use the most robust data first. If linking GP data, get primary care involved early on and ensure you take this to local medical committees.
- Be realistic. Link a small number of datasets initially to prove the benefits of the data linkage concept, and then expand from there. Avoid the temptation to get overambitious and link too many datasets together.
- While there is a great deal of mandatory standardisation of data, data quality could be improved everywhere. Be careful when using new sources. Make good connections with each data supplier and understand your confidence level on each field.

[17] A data dictionary is a collection of the names, definitions and attributes for data elements and models in a database. It contains information such as data ownership and data relationships and helps to organise data and prevent data redundancy issues.

[18] Deterministic linking uses a unique identifier such as an NHS number to link records. If the unique identifiers are unavailable or of poor quality, probabilistic linking may be used instead. This approach looks at several identifiers in combination to identify records in different datasets that have a high probability of belonging to the same person/organisation.

- Provide an API (or ‘widget’) to allow partner organisations to submit data to the linked dataset that is automatically pseudonymised.
- Data can be de-personalised and made available to analysts in care providers to enable them to compare it with their local system and flag any discrepancies.
- Compare trends over time to check data quality; if a figure is relatively consistent and then suddenly changes, investigate the cause.
- AI methods are available which could be explored further to improve data quality and data matching. Click here to find out more.
- Where actions are needed to improve data quality issues unearthed by linkage, make sure that those capturing data are involved with the ‘why’ and ‘how’. Staff responsible for capturing data as part of their business-as-usual activity will need to ensure complete, accurate and up-to-date information is being captured in IT systems.

Using the data

Step 7: Data validation and strategic analysis

- Work collaboratively with clinicians and public health consultants who are specialised in PHM.
- Talk to data suppliers/in-house teams to make sure the data is as well populated as possible, and carry out extensive testing.
- Ensure you investigate and address any gaps in the data that could affect future analyses.
- Be clear on your data storage and processing boundaries; keep only bespoke data extracts for the particular piece of work the data relates to. This will help give people confidence that you are not building something that could be misused in future.
- Maintain organisational hierarchies and lookup tables in the database. Hierarchies will order groupings of attributes to reflect their relationship with other attributes. Lookup tables will identify a secondary value based on a primary value (i.e. will retrieve values from related tables).
- Make sure you have a robust update process for new data coming in that allows for removal of duplication or updating a record.
- Look to publish findings in the open unless there is a justifiable reason not to do so. Share methods and code. This could be done via Future NHS or by setting up an MS Team for all national PHM/data linkage teams to share best practice, for example. Put your matching statistics ahead of the results when publishing outcomes.
- Get regular feedback and insight from people who are using the linked data and from the people who are recording the data; both can have different ideas about quality.
- Consider having an independent assurance function to check that people are doing the right thing, and be prepared to call out incorrect behaviours.

Appendix 2: Datasets

There are numerous datasets which could be included in data linkage, including:

- Real-time health and care data:
    - SUS
    - GP
    - Community services
    - Mental health
    - Clinical data – diagnostics, bloods, imaging etc.
    - Social care providers, local authority social care management
    - Office for National Statistics
    - NHS demographics – Spine (PDS)/NHAIS
    - NHS 111
    - Out of hours data
    - Ambulance trusts and patient transport
    - Hospice data
    - Workforce data – including, for example, who treated this patient and how that impacts continuity of care
    - Patient-reported/collected outcomes – PAM, WEMWBS etc.
- Wider determinants
    - Education/environment
        - Education (with the possibility of linking/aligning pupil reference number as person-level identifier)
        - Police and criminal justice system
        - Deprivation index
        - Weather conditions (including air quality (pollution/pollen count), key for those with respiratory issues)
        - Transport, GPS/accelerometry
        - Census data
    - Finance
        - Consumer data (e.g. banking/financial, telephony, energy consumption data)
        - Employment
        - DWP data
        - Council tax
        - Food bank usage
    - Local authority and other local services
        - Fire service
        - Benefits and housing
        - Citizens Advice Bureau
        - Leisure memberships

- Assisted bin collections
- Loneliness services
- Children’s social care
- Troubled families/social services
- Public health teams based in unitary authorities have commissioner and provider arms; the latter will be collecting data on health improvement services they commission, such as smoking cessation, alcohol and substance misuse
- Third/charity sector providers (some available via DSCRO, such as social prescribing)