Public Health Research Data Forum

Enabling Data Linkage to Maximise the Value of Public Health Research Data: full report

March 2015

wellcome trust

Enabling data linkage to maximise the value of public health research data

Final report to The Wellcome Trust

Acknowledgements

This research was undertaken by a project team, whose membership was as follows:

University of the West of England, United Kingdom (Faculty of Business and Law)

- Elizabeth Green , Research Associate, Bristol Business School
- Felix Ritchie, Associate Professor of Applied Economics, Bristol Business School
- Don Webber, Professor of Applied Economics, Bristol Business School

University of the West of England, United Kingdom (Faculty of Health and Applied Sciences)

- Julie Mytton, Associate Professor in Child Health, Centre for Child and Adolescent Health
- Toity Deave, Associate Professor in Family and Child Health, Centre for Child and Adolescent Health

University of Cape Town, South Africa

- Alex Montgomery, Research Associate, DataFirst
- Lynn Woolfrey, Manager, DataFirst

Centre for Injury Prevention Research Bangladesh

- Kamran ul-Baset, Senior Scientist, CIPRB and Associate Director, RTi research centre
- Salim Chowdhury, Director of Training and Education, CIPRB, and PhD fellow at the Department of Public Health Sciences, Karolinska Institutet

The project benefited greatly from the support and advice of Jane Simmonds and David Carr at the Wellcome Trust, particularly in the reporting stages.

We are grateful to those who agreed to be interviewed for this report, formally and informally. The views and opinions expressed in this report are those of the authors, and do not necessarily represent the views of the Wellcome Trust, the Public Health Research Data Forum, or any other body. Errors and omissions in the interpretation of research findings or interviews remain the responsibility of the authors.

Contents

Executive Summary

This project was commissioned by The Wellcome Trust on behalf of the Public Health Research Data Forum. The project aimed to identify the gains to public health research from linking existing data sources, the opportunities in and barriers to such data linking, and how the barriers could be overcome. The objective was to deliver a set of practical recommendations for realising the gains from data linkage. 'Data linkage' is broadly taken to be the linkage of health data within and across organisations, and as well as linkage between different data sources such as hospital admissions, cancer registries, and socio-economic surveys.

The research strategy was to use a mix of literature review, case studies of data linkage projects, and interviews with selected individuals involved with data linkage. The study looked at low-, middle- and high-income countries to ensure that lessons learned would have wide applicability. Barriers to useful data linkage were analysed from statistical, operational and institutional perspectives. Given the vast amount of information on data linkage theory and practice, this project focused on useful illustrative examples as opposed to an exhaustive review of the field.

Findings

A key concern was the issue of whether narrow informed consent (NIC) should be the primary basis for research. Many researchers observed that broad consent was practical and acceptable to the public when data was collected for research purposes. However, in public health much of the value comes from linking data collected for administrative or statistical purposes, for which it was not practical to obtain consent. Even where it was practical, the statistical consequences of insisting on NIC severely damaged the potential in the data. Hence, there was universal agreement amongst respondents that a practical exemption from NIC for statistical research was essential for high-quality high-benefit public health studies. As such the forthcoming EU Data Protection Regulation was causing great concern amongst the European interviewees. There was an equally strong common understanding that the quid pro quo for a research exemption was an appropriate social contract, where clear objectives and accountable, transparent processes provide the guarantees that the public needs about the use of their data.

It was recognised, particularly but not exclusively by members of the research community, that there is a need to change the tone of the debate: from the assumption that nothing can be released unless it is explicitly allowed, to a position where all data can potentially be used unless it can be shown to be unlawful, unethical, or unachievable in a manner which protects confidentiality. Although a small point, this change in perspective has a major impact on the type of discussions to be had. A related development is the growing fondness for principles-based planning. Both of these seek to put the objectives of data access and linkage at the forefront of decision-making.

Related to this was the perception that decisions are often made without sufficient reference to evidence. This was particularly the case when considering how research access to sensitive data was managed. Those involved in the design and management of research facilities or pathways saw this as a 'solved' problem: different implementations over many years showed that, despite theoretical concerns, in practice this was a very low risk activity. However, because this was seen as unremarkable by this community, this may not have been communicated well enough to external interested parties such as legislators or data depositors. Hence, the data management community

may have inadvertently created a climate where research data access is viewed as high risk and difficult to manage.

Key to the successful operation of any linked data project is the relationship with others: the public, the researchers, the data depositors, the research ethics committee (REC). The public generally are very supportive of health research (although this is sensitive to the framing of questions), something which we may not acknowledge enough. Public support for research is closely related to the trust in the institutions: the public are comfortable with broad consent, which implies trust that the organisation asking for consent will 'do the right thing'. Health organisations tend to do well as 'trusted' bodies, often being viewed as among the most trusted.

Relationships with data depositors and RECs can make the difference between a successful project and an administrative nightmare. For high-income countries (HICs), strong organisational links seem to make the difference with data depositors, whereas for low- and middle-income countries (LMICs) personal links seem to matter more. In LMICs, the level of association with governments can also prove important, as there may be a higher risk of being linked to the ideals of a particular regime rather than working for the public good.

It was widely noted that researchers can be part of the problem – they may be unwilling to make data accessible, even though funders require it. This is understandable – researchers might have spent many years developing data resources, and, as noted below, such efforts are not always rewarded in funding or publications. This may be something that funders are best placed to tackle.

Data quality is a major issue for LMICs, whereas it seems of much lower importance for HICs: it exists as a practical problem (particularly in terms of accuracy), but the institutional barriers are what mostly exercise research data managers. The opposite seems to be the case for concerns over slow processes and the perceived waste of resources in getting agreements: these are highlighted in HICs, but are much less frequently raised in by LMIC respondents. However, it could also be a question of more realistic expectations in LMICs.

The case of South Africa seems to suggest that there is a natural progression from operational problems to more statistical ones as data linking increases and becomes more the norm. Given the longer experience of HICs in data linking and managing, there may be gains to be made from sharing information about skills, data facilities and storage models, allowing LMICs to avoid some of the problems experienced by HICs.

The broad conclusions of the report can be summarised as follows:

- Theoretical or statistical challenges for data linkage can generally be seen as solved, at least for practical purposes.
- Practical issues still exist, and are much more important in LMICs where data quality is lower:
    - Good consistent identifiers substantially improve outcomes, but should not be pursued at the expense of the variables of interest.
- There is a need to ensure that decisions about linkage are well-informed and evidence-based:

    - Narrow informed consent alone is not a basis for good epidemiological research; some form of workable research exemption is necessary.
    - There is ample evidence to show that the social contract can be managed effectively.
    - There are substantial differences in the ethical positions taken by those in authority, which seem more to do with cultural or institutional factors than genuine ethical matters; this variation in practice (even within countries) has a substantial negative effect on research.
- The general public (at least in HICs) is very supportive of using linked data for research:
    - Trust in institutions is one of the most important factors for public acceptability of research use of data, at all levels of decision making.
    - Trust is fragile - one high-profile incident could set research data access back a long way, but memories are short.
    - The framing of questions is crucial to issues of public acceptability.
- The data management community largely views research use of data as relatively low risk, which can managed safely and effectively:
    - For this community, safe management of data is a practical matter of designing systems, procedures and training.
    - This view (and the evidence base) does not seem to be communicated well outside that community, who are more likely to focus on theoretical risks.
- Cultural issues are important in determining the success of a project:
    - Personal relationships and personal authority can go a long way to resolving (or creating) problems.
    - Turf wars and power relationships can create reasons for excessive regulation.
    - Some academics are resistant to sharing data, even where funders require it.
    - This was identified as a more significant barrier in HICs, compared to LMICs
- Incentives to manage and link data are weak:
    - There are few incentives to specialise or develop expertise in data, per se.
    - Transferring knowledge to LMICs is a resource-intensive process.
    - Data linking is a long process which should be better viewed as an investment in a cumulative store of knowledge.

Recommendations

Our recommendations to the Public Health Research Data Forum (PHRDF) are largely concerned with distributing useful and accurate information to change ideas about data linkage and show the possibilities to interested parties. We believe that a common perspective from a critical mass of funders would substantially improve the environment for and practice of data linking.

Our recommendations are grouped around two topics: setting the conceptual framework, and finding solutions to practical problems.

Set the conceptual framework to control the debate

The aim of this set of recommendations is to change the general language of debate to make it more supportive of data linking, and provide the conceptual basis for strategic thinking on improved data access.

- Change the language used when discussing data access from default-closed to default-open
- Develop and promote high-level principles for research access to data and data linking
- Encourage practitioners to share their knowledge and experience of effective risk management in research access
- Develop a toolkit of coherent cases, backed by evidence, which can be used for advocacy purposes in policy discussions
- Produce guidance on best practice ethics processes which encourages collaboration and co-operation

Help resolve practical problems with specific advice on good practice which seems to work

There are also a series of practical steps through which funders could support researchers in developing data linkage activities.

- Encourage the use of remote technology to allow knowledge transfer between HICs and LMICs, particularly collaborative working tools
- Provide dedicated funding for the creation and management of data resources as a distinct element in research grants
- Invest in PhDs as a cost-effective long-term investment to develop data expertise in LMIC and HIC settings
- Draft guidelines for research teams on addressing practical issues in enabling data access and linkage
- Build up a record of ‘useful’ precedents, experience and exemplars

Part I: Background to the project

1. Introduction

In recent years, increased use of existing data resources through improved access arrangements and data linkage has come to be seen as one of the most cost-effective ways of supporting research in public health and epidemiology[1]. Re-using and extending existing data from primary data collection has the advantages of immediacy and increasing the return on the investment in data collection. Using administrative data for research can be more complicated, but this is offset by the depth of coverage available from such data.

Public health research needs to consider the wider social determinants of health. Traditionally these determinants may have been the separate preserves of social scientists and health researchers, each group working to collect data and analyse their own data; but inter-disciplinary studies and data linking is increasingly seen as the norm.

This project was commissioned by The Wellcome Trust on behalf of the Public Health Research Data Forum. It aimed to identify the gains to public health research from linking existing data sources, the barriers to such data linking, and how the barriers could be overcome. The objective was to deliver a set of practical recommendations for realising the gains from data linkage, by a mix of literature review and interviews with relevant experts.

This report is structured as follows. In the remainder of Part I, we summarise the research strategy for the project, including the initial twelve questions asked by Wellcome Trust in the project brief. We then provide a summary of the current literature on data linking: what it is, its value to public health research, and the problems associated with it.

Part II summarises the findings from the study. It breaks these down into four aspects: conceptual issues, the environment in which linking takes place, practical matters, and future concerns. This section brings together information identified from a literature review, a series of case studies and the views of those interviewed for this report.

Part III discusses the findings in the context of the original twelve questions from The Wellcome Trust. It then makes practical recommendations based upon study findings.

Annexes

This report assumes some familiarity with the terms and concepts of data linkage and data matching. For those unfamiliar with the subject, Annex A (part 1) contains a brief summary of the technical aspects of the topic. Annex A (parts 2 and 3) contain the extended discussion of the literature summarised in this section.

Annex B describes the research strategy in more detail and lists the named interviewees.

Annex C contains a set of eleven case studies. These detail particular examples of data linking projects which have useful lessons. The case studies are intended to be illustrative, not exhaustive.

[1] Epidemiology (statistical analysis of the population) is an important component of public health, but the latter also includes clinical trials and matters of health care provision. As this report focuses on the analysis of secondary data, we treat public health research and epidemiology as broad synonymous for simplicity.

Note on definitions and abbreviations

In discussing data ownership and use, terminology is important. For example, what does the term 'data owner' mean? It may just refer to the organisation that decides how data are to be used. However, an objection is that the *right* to decide how to use that data should not be vested in the current (temporary) holder of the data, as the right to manage that data ultimately should lie with the person that the data refer to. Moreover, it could be argued that calling an organisation a 'data owner' encourages it to start thinking in terms of '*my* data' rather than 'data which has been provided to me'.

Other terms are problematic as well: for example, 'data controller' has a specific legal meaning in much of Europe, 'data guardian' a specific interpretation in terms of health data in the UK, and so on.

To provide consistency throughout the report, the following terms are used, without reference to any specific legal or ethical role:

Data subject	The legal or natural person that the data refer to
Data collector	The body that acquires the data about a subject
Data depositor	The body that deposits collected data with third parties for re-use
Data linker	The body responsible for linking data sources together
Research data manager	The manager of data made available to researchers
Data user or researcher	The end user of the data for research purposes

Note that an organisation may embody more than one of these roles.

The following **abbreviations** are used in this report:

EC	European Commission
EU	European Union
GP	General practitioner
HIC	High-income country
IC	Informed consent
LMIC	Low- and middle-income countries
NHS	National Health Service (UK)
NHSCIC	National Health and Social Care Information Centre (UK)
NIC	Narrow informed consent
PCP	Primary care provider
PHRDF	Public Health Research Data Forum
REC	Research Ethics Committee
SDC	Statistical disclosure control
TTP	Trusted third party

2. Project strategy

2.1 Research questions

Twelve questions were identified in the initial project tender:

- What are the potential benefits (including impact) and opportunities of linking research data (now and in the future)?
- What are major challenges (e.g. technical, ethical, legal, financial, cultural) that prevent these benefits being realised and how might those challenges be addressed?
- What specific challenges exist in relation to: linking, harmonising and pooling data across national boundaries?
- How is effective data linkage defined conceptually and in practice?
- What lessons can we learn (for research funders, researchers, policy makers and health practitioners) from existing data linkage initiatives in terms of the systems that they are using and the training that they are providing?
- What best practice principles should be adopted and what practical solutions could be considered?
- What is the relative position of different fields in relation to utilising data linkage (e.g. biomedical, health, economic, environmental, social data) to produce evidence to support policy and delivery of health services and medical interventions?

For low- and middle-income countries:

- Do the challenges and benefits differ: (i) within and between low and middle income countries (LMICs) and (ii) between LMIC and non LMIC countries? Are there transferrable lessons?
- Are there specific approaches that have been effective in LMIC and non LMIC countries?

Future trends:

- What are the new and emerging data sources which have the most potential in relation to data linkage in the field of public health?
- What are the new and emerging technologies and methods that are having an impact now or in the future on data linkage in the field of public health? What are the implications for governance?
- Where might we be in ten years from now in terms of data linkage?

These questions will be considered in the conclusions section as a way of bringing together the findings of the report.

2.2 Research strategy

The research strategy was to use a mix of literature review, case studies of data linkage projects, and interviews with selected individuals involved with data linkage. The study looked at low-, middle- and high-income countries to ensure that lessons learned would have wide applicability. Barriers to useful data linkage were analysed from statistical, operational and institutional perspectives. Given

the very large amount of information on data linkage theory and practice, the project focused on useful illustrative examples rather than an exhaustive survey of the field.

Data collection had three elements:

- a review of the academic and non-academic literature, based on online searches;
- case studies of particular data linking projects;
- interviews with experts involved with data linking.

The review of literature used a combination of web-based searches and the personal experience of the team in data access, public health and cohort studies. See Annex B for further details.

The case studies were chosen on the basis of the literature review, and the personal knowledge of the team. The case studies generally led to interviews with involved parties. Because of limited resources the decision was taken to concentrate on a variety of experience, and so not all of the case studies initially identified were followed up; for example, the UK has many cohort studies, but ALSPAC was taken as a representative example.

The interviewees were chosen by a mix of convenience sampling ('easily available' interviewees) and snowballing (one interviewee leading to other interviewees). The initial selection of interviewees was driven by the personal knowledge of the teams in the UK, Bangladesh, South Africa, and Sweden. Contacts were also provided by Public Health Research Data Forum members.

In addition, it became apparent that ethics committees played an important role in the success of data linkage, and so it was decided to interview a small number of ethics committee members. The study was also informed by attendance and discussions at the 2015 Computers, Privacy and Data Protection conference in Brussels, January 2015[2].

Because the number of interviewees was small and identifiable, no direct sourced quotes are used in the main body of the document; this enabled participants to speak more freely. As well as the formal interviews, a number of informal face-to-face and telephone discussions took place. The case studies are sourced to a particular person, and have been checked by the relevant interviewee; however, they still reflect the personal opinion of the individual and should not be taken as the official position on data linkage of any institution or organisation.

2.3 Project team

The project team consisted of researchers from the UK, South Africa, and Bangladesh. The project was managed by the University of the West of England, Bristol. African contacts, interviews and insight were organised and carried out by DataFirst at the University of Cape Town. Interviews and perspectives on Bangladeshi experience were organised and carried out by the Centre for Injury Prevention Research Bangladesh in Lahore. All other interviews were carried out by the UWE team.

[2] http://www.cpdpconferences.org/

3. Data linking in literature

This section provides a very brief overview of data linking from the literature. A more detailed review, with references, can be found at Annex A.

3.1 Basics of data linking

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.

3.1.1 Identifiers and identification

When linking data, variables are typically split into:

- Identifying variables (for example, name, address, medical insurance number);
- Variables of interest (age, gender, income, illness, occupation etc.)

Direct identifiers (such as name and address) allow individuals to be identified exactly. Indirect identifiers only identify individuals in combination with other information.

Direct identifiers are typically of little interest to researchers; their value is in allowing the data to be linked, and so they are removed from datasets before research access is allowed. Indirect identifiers and variables of interest often overlap; for example, age, gender and ethnicity can be used to identify an individual but are also typically valuable explanatory factors. Hence, a useful dataset is likely to have some characteristics which will allow the individual to be-re-identified from the data, even if this is very unlikely; this is called 'pseudonymised" (pseudo-anonymised) data.

3.1.2 Types of data linking

A number of techniques are available for data linking.

Exact/deterministic linking

Exact (or deterministic) linking is possible where a unique identifier is shared between two data sources. For example, in the UK, a National Health Service (NHS) number is used to link data across NHS medical records. The obvious advantage of exact matching is that the link is certain and simple to effect. A secondary advantage is that the match field is typically a non-informative reference number, and so there is less concern about identifying information being released through accidental exposure.

Exact matching requires that the match field is unique and accurate; this is most likely to occur in well-resourced administrative systems which have a substantial benefit from a common reference number. It is less successful when trying to match, for example, names and addresses open to mis-entry ("John Smith" in one dataset appearing as "J Smith" in another). In these circumstances, an alternative approach is probabilistic matching.

Probabilistic matching

Probabilistic data matching is a well-established and common solution for data linkage. This compares the identifying variables across two or more datasets to estimate the probability that two records relate to the same person. This method explicitly acknowledges that data might be

inaccurate, incomplete or entered differently in data sources, and so it is more general than exact matching.

As this is an estimate of how likely it is that two records refer to the same person, there is the possibility of both *false negatives* (a true match not being recognised) and *false positives* (declaring two records to refer to the same person when they do not). Reducing the chances of one increases the chance of the other, and so the way that a probabilistic match is set up reflects the preferences of the person doing the linking. Matching software usually allows the linker to specify the expected ratio of false positive/negative readings. Automatic matching is often supplanted by 'clerical' matching (a human looking at the records to improve the match rate).

Preparing the data for linking can require a substantial amount of data cleaning, and the matching of fields can also require extensive computational resources. Nevertheless, this is a tried-and-tested method which is, by design, more tolerant of data errors then exact matching.

Statistical linking and data fusion

Statistical techniques (sometimes called data fusion) have been developed to allow analysis where the records of two different individuals have been linked as if they refer to the same person. This has been exploited by commercial organisations as a way of generating synthetic data for analysis. In public health, its main purpose is to allow one to build simulation models for policy evaluation.

Multilevel linking

Data linking need not be at the level of personal records. As noted above, there can be substantial gains from linking personal data with, for example, environmental data. The match is also typically exact (one knows, for example, the area the subject comes from) and has lower (but not always negligible) confidentiality risks.

3.1.3 Characteristics of types of data

Cross-sectional survey data

Surveys tend to be used to collect socio-economic data; as they are designed for statistical purposes, they tend to be high quality. The major concern is ensuring that the data are representative, as most data are collected as samples from the population of interest. There is also less opportunity to carry out validation checks using data from other sources. Finally, because the data are typically pseudonymised, linking can be problematic.

Cohort studies and longitudinal studies

In cohort studies the subject is repeatedly interviewed, and the cohort planners will actively try to ensure that contact is maintained with the respondents. This provides additional checks for the quality of the data, as well as a mechanism for following up queries. Cohort studies have many advantageous statistical properties; their major drawback is the cost associated with managing a complex data collection operation over a long period, and the loss to follow up of participants.

As far as linkage is concerned, cohort studies should be an easier proposition than cross-sectional studies as maintaining accurate identifying information is essential to keep the cohort going. Linkage can also pay dividends to the cohort, by feeding back information for future studies on the cohort.

Register data

Many countries have population registers; some are general – for example, to manage ID card systems – but others may be specific to particular areas, such as cancer incidence. These have great statistical potential: they reduce the problem of individuals being selected into surveys or cohorts in a biased manner, and produce ready-made control and treatment groups.

In some countries registers have common personal identification numbers, making linkage fast and accurate. Even if different IDs are used, registers are designed to be continually updated with new information, and so linkage is facilitated. The most extensive systems of general registers occur in the Nordic countries.

Other administrative data

Administrative data (that is, data collected through normal operations) can often be a census of the population of interest. Hence, as for register data, administrative data can be used to reduce selection bias and provide control and treatment groups.

Administrative data is collected for operational needs, not statistical ones. Hence the data may suffer from quality issues, semantic problems (administrators understanding of the data may differ from what the researcher wants), and a choice of variables limited to business needs of the data collector.

3.1.4 Confidentiality issues

Use of sensitive data for research causes concerns about whether that data is being used safely. Two scenarios that are typically used are an accidental risk of release of confidential data (for example, by someone leaving a CD on a bus), and a researcher deliberately trying to identify someone from the data.

The accidental-release scenario does occur, although it is extremely rare and the impact low (particularly when compared to releases of data from administrative sources, for example). The deliberate-release scenario, despite its popularity in the statistical literature, has almost no evidence to support it, at least in the last fifty years.

The excellent security record of academic researchers is largely down to two factors. First, a research dataset is easier to manage and control than an administrative data set, where many people may have access to fully identified records. Second, research data is almost always de-identified as soon as is practical, and so the potential risk of loss data is small. Third, most researchers go through extensive data protection training, and data management plans are a typical component of ethical committee approval. Finally, the research community has developed a range of technical solutions allowing data of different levels of sensitivity to be managed in a variety of computing environments.

Data linking raises more concerns, as the fact that the data is to be linked means that identifiable data is more likely to be shared. In practice, however, this does not significantly increase risk as the best practice models that most facilities adhere to ensure that the linking is kept a separate process from the delivery of research datasets.

Hence, while the use of sensitive data for research does create a confidentiality risk – and linked data have an increased risk – the empirical evidence suggests that this is an extremely low risk which can be managed effectively.

3.2 The value of data linking

The ability to link different data sources together is crucial to epidemiology for a number of reasons, as summarised below. For a fuller discussion, see Annex A.

Treatment and control groups

Many statistical procedures require the identification of 'treatment' and 'control' groups (that is, those who have and have not gone through some experience or treatment). Often a single data source will just have one or the other, and so linking makes this essential technique possible.

Range of topics

Combining clinical data with other data sources may allow the data to be broken down in different ways, and make it possible to answer questions which a single data set cannot resolve.

Long term study

Health events can be experienced over an extended period, and tracking all relevant events over such a long period may not be feasible in a single database without excessive intrusion and/or cost. Using additional data which records such information as a matter of course can improve the accuracy of data collection and reduce the burden on both observer and subject.

Retrospective analysis

Some conditions may not manifest themselves until many years after the initial incidence; alternatively, an illness may appear quickly but have contributory factors going back far into the patient's past. In both these cases, to study the illness it is necessary to use historical information which was collected for other purposes, such as administrative data, vital events data, civil registration data or other sources. Such information is particularly valuable in the case of rare health events, where it is difficult to identify in advance who might be susceptible to illness.

Prospective data collection

A parallel to the retrospective study is the prospective cohort study, identifying a cohort of people and following them over time, in more or less detail. As for retrospective analysis, the great statistical advantage is that groups are chosen before any medical conditions arise, and so 'baseline' information on all subjects can be collected before treatment and control groups are identified; again, data are collected throughout the period and so recall error is not an issue.

Co-morbidity

Multiple health events can occur at the same time, or be associated with multiple concurrent socio-economic factors. These might not be recorded together as each data collection agency is focused on the most relevant condition. Bringing these records together allows co-morbidity to be investigated.

Checking and improving data quality

All data contain errors to a greater or lesser degree. Combining multiple datasets allows the consistency of data to be checked, and potentially gaps to be filled in.

Analysing rare events

By their nature, it is difficult to generate sufficient information on rare events from single data sources; but pooling data from different years and data sources (perhaps even different countries) can generate sufficient data to model these rare events.

Linking personal data to the environment

By combining personal data with information about groups, areas, systems and so on, it is possible to draw out contributory factors which reflect structures in society, such as proximity of health care facilities to particular groups.

Generating useful tools

Linking data from multiple sources can allow population level tools to be developed, such as simulation models.

Making data analysis more timely

Linking data from existing sources for analysis may well be the quickest way to get the answer to a statistical problem; there is no additional time to collect the data, and so analysis can be achieved relatively swiftly.

Generating cost savings

Dedicated data collection is expensive, particularly from medical sources. If that data can be re-used then the public benefit can be substantial.

Enabling International comparisons

Sharing or linking data or results between countries allows the effect of national environments to be studied; and it may be necessary to boost study numbers in very rare illnesses.

Delivering Interdisciplinary research benefits

Epidemiology explicitly recognises that the health of the public can be determined by socio-economic factors as well as by viruses or bacteria, and so an inter-disciplinary research environment might be more successful at identifying causes and effects.

3.3 Problems with data linking

In theory, a researcher wanting to link data sources can call on many statistical and practical resources. In practice, data linking is much less straightforward. While the conceptual environment is well established, the practical difficulties faced can be substantial. This short summary review is organised around three topics: statistical issues, operational/technical issues, and institutional issues. A more detailed discussion can be found in Annex A

3.3.1 Statistical issues

Whilst all research data has some limitations, linking data generates a specific additional set of problems.

When analysing a single dataset, some measurement error can be tolerated, but this can substantially affect successful link rates. Similarly, small variations in consistency can have a disproportionate effect on match quality.

Linking data from two samples is likely to substantially reduce the amount of usable data. On the other hand, where one or other dataset is a census (for example, a register of all diabetes patients), linking enhances the utility of the matched data. A related issue is how well the statistical characteristics of the match data are known; again, when at least one file is a census this problem is simplified, but if two data sources are sampled, little is known about the characteristics of a matched dataset if the assumptions about the data do not hold true.

Broadly, however, while there is statistical research going on, the theory of data linking is settled, robust, and uncontroversial; and there are off-the shelf solutions to implement these methods.

3.3.2 Technical and operational aspects of data linking

There are five stages from proposing a project to getting linked data used in research:

- Acquiring permission to link;
- Agreeing the hosting protocol;
- Acquiring the data;
- Providing access to researchers;
- Using linked data in research.

The first three stages are complicated by the need to obtain agreement from multiple organisations. These organisations may differ in their interests and objectives, their perspectives on security, the actual and perceived risk associated with use of their data, and their understanding of the research environment. Getting agreement from data depositors is therefore more complicated than for single datasets, although this is more a question of degree rather than substance.

Of more concern is that the need to link data means that one organisation will probably need access to identified data from at least one other another organisation (in contrast, for single-source analysis, only the original data collector needs to see data with detailed identifiers). The research industry has largely solved this problem by the use of 'third party linking', where one organisation is given the identifying data only (not variables of interest) and is charged with creating an anonymous link field which can replace the identifiers on the source datasets; these can then be linked through exact anonymous matching.

Third parties can either be 'trusted' or 'untrusted': in the former case, the third party receives the original identifiers, while in the latter it gets identification information transformed to be uninformative about the data subject. Untrusted matching is unappealing from statistical and operational perspectives, whereas 'trusted third parties' (TTPs) are straightforward to implement, are better able to deal with data problems, and have a good record of managing data confidentiality. Hence TTPs are widely used, familiar, and relatively uncontroversial in practice.

Similarly, providing researchers with access to confidential data can be seen as a question of picking the right off-the-shelf solution. There are a wide variety of technical and managerial approaches available, and there are frameworks to help research managers decide which solution is most appropriate (and convince the data depositors that this choice is correct).

Linked data does present extra challenges when considering the research use of this data. The data may be more complicated than single source data: it may be from different time periods, from

different types of data, and with different sample characteristics. However, the main problem is that none of the data depositors is familiar with the full dataset, each having only contributed a part of it. One solution is to make the research manager the point of expertise in the data (rather than the data depositors); this can have additional advantages in terms of improving the engagement with researchers, but it does have cost implications.

In summary, the operational aspects of data linking, while often complex, do not present major unsolved problems. Many of the issues are similar to those experienced with single-source data, and the aspects specific to linked data (circulation of identified data, user support) have tried and familiar solutions.

3.3.3 Institutional aspects of data linking

Much of the literature is concerned with institutional aspects of data linking. Unlike the statistical and operational issues, this literature contains a number of unresolved debates.

Legal issues

Consent

A person consenting for his or her confidential data to be linked and analysed is often referred to as the 'gold standard' gateway. It provides both an ethical and a legal framework for managing and using data. However, there are a number of practical, ethical and statistical problems:

- it may be difficult or impractical to contact the subject;
- consent may lead to biased samples if those giving consent differ from those refusing it;
- use of data may identify family members, for example in DNA samples;
- gaining consent may be undesirable as it breaches confidentiality (for example, by revealing selection criteria).

It is straightforward to show that using only data for which consent has been given can lead to significantly biased outcomes. The impracticality of gaining consent, for example for linking to historical data in retrospective analyses, can also dissipate any cost advantage from observational studies. Hence, many researcher analyses argue that consent is desirable but that statistical demands need to be taken into consideration, and the feasibility of non-consensual gateways explored.

It is also not clear what is meant by 'consent'. 'Narrow informed consent' (NIC) to a very specific project may satisfy stringent ethical concerns, but may overly limit research. Broad consent (BC), where the subject agrees to their data being used in unspecified ways but by a trusted body, is much more common; indeed, it is necessary for cohort studies and other long term analyses where the use of data is unknown at the beginning. However, there are concerns that too broad a consent is not consent at all.

There is also debate about whether consent should be 'opt-in' (no participation unless explicitly agreed) or 'opt-out' (participation is assumed unless the subject chooses not to take part). The choice has been shown to significantly affect participation rates, but it also raises questions about whether the consent is truly 'freely given'.

Because consent is not problem-free, many countries have a legally mandated gateway allowing access to data for research purposes, a 'research exemption'. Such legislation typically also specifies that there be appropriate checks and balances to ensure that data collection is consistent with the spirit as well as the letter of the law. However, the use of research exemptions is not uncontroversial, as the ethical basis is subject to challenge (see below). Perhaps more importantly, government-mandated use of personal data without consent can be perceived as 'Big Brotherly'.

Competing jurisdictions

Even if the legal framework is clearly defined, projects may suffer from needing the approval of multiple jurisdictions. This can be seen as a failure to distinguish between legal responsibility (to carry out due diligence on potential projects) and between gathering evidence (by accepting, for example, that another ethics committee is competent to carry out due diligence and so take the decisions of that committee as evidence of compliance). Given that some of the most interesting developments in public health are the relationship between medical and socioeconomic factors, competing jurisdictions for approval are likely to be a concern.

Law versus custom

Law is rarely a black-and-white issue; it needs interpretation in particular cases. However, most researchers are not specialists in law, and it is common for custom to be seen, over time, as law. This is most likely to occur where, in the absence of explicit legal statements, institutions are tasked with deciding the interpretation of the legal framework. Hence, research gateways can suffer from 'regulatory capture' by institutions keen to ensure that their interpretation of law prevails.

Defining confidentiality

Whilst legislation may use such terms as 'confidential' and 'anonymised', there is no legal definition. Instead it may be left open for a competent authority to determine, and/or reference to be taken to 'reasonableness'. Hence, a key part of the legal framework is left open to human interpretation, and two organisations considering the confidentiality of a linked data source can come to different conclusions, each consistent with the data depositor's perspective. This complicates any discussion on appropriate technical solutions.

Ethical concerns

Ethical assessment requires balancing competing subjective claims: the rights of the individual against the rights of society. Putting aside the statistical issues associated with consent, the standard starting points are that:

- the individual has a right to privacy and therefore control over his or her data (i.e. informed consent must be present);
- the government has a duty to act in the interests of society as a whole and may override the wishes of an individual (i.e. informed consent cannot be insisted upon).

The first point is found in numerous documents where the need to avoid harm is emphasised. However, the argument that consent is neither necessary nor sufficient to prevent harm is easily demonstrated.

The case for a research exemption has been made as a 'paternalist' argument: the state knows what is best. However, as this can be used to justify a range of dubious behaviours, the case is more often made in the form of a 'social contract'.

The basis of this social contract is 'reciprocity' (sometimes 'solidarity'): research is uncertain and so there is no direct connection between costs (allowing data to be used in a study) and benefits (the findings of public health research). I am part of society; I help with 'research' without knowing who I help, because I expect others to help me without knowing it. This produces a moral argument for participation in research, and a rationale for the state to over-ride particular preferences: to prevent free-riding.

A third argument is based on self-interest: although research is uncertain, participation can help the development of treatments beneficial to the subject. The loss of privacy is small and manageable and the potential gains large, if uncertain. This argument is often used to persuade participants to give consent to their data being used for research. However, it is less clear that it provides a rationale for a research exemption: the obvious problem of this cost-benefit argument is that sometimes the cost definitely exceeds the potential benefit: for example, taking tissue samples from elderly men to study childhood diseases or ovarian cancer.

Within the public health profession there is therefore a broad consensus: in principle, public interest can be allowed to take precedence over individual consent where the statistical needs and public benefit justify it. Note that this does not say what should happen in a particular case; the key issue is that a research exemption of some form is necessary.

As it stands this is not controversial, but it can be seen as leading back to the 'paternalist' argument, and so most authors accept that reviewing the balance of public and private costs is an essential part of research approval.

Much of the linked-data literature concentrates on the use of administrative data. Unlike statistical data collection, the primary ethos of administrative data is to serve the customer. Hence, for example, a GP may consider that doctor-patient confidentiality is his or her primary responsibility, not supporting the health service's research programme.

Cultural barriers

Public attitudes to data sharing

Public expectations can have a profound effect on the prospects for research use of confidential data. Linking datasets can be more problematic in the public's eye because it immediately brings to mind the image of a government actively trying to find out more than the individual is prepared to disclose. In theory, gaining consent rather than using research gateways in legislation can legitimise linkage in the public eye, but, as was noted above, consent may not always be desirable or feasible.

Studies tend to show that the public is comfortable with:

- their data being used in research, particularly by academics;
- their data being made available to 'trusted' organisations;
- broad consent being used to carry out studies, and no need to obtain consent for specific projects;

- health data being linked with other data to carry out research.

The most important seems to be the second: if an organisations is 'trusted' to look after one's data, then all of the others tend to follow. Usefully for the purposes of this report, health organisations have repeatedly been found to be among the most trusted types of organisation.

However, while these general findings seem to be robust, they are sensitive to the way questions are framed, as well as, for example, media stories about privacy and data security. The general public find it hard to follow, understandably, quite complex issues such as anonymisation and data flow models; they react with concern to complexity. People tend to look more favourably on things they have personal experience of, and rely upon media reports for more abstract concepts. The answers therefore depend upon both the framing of questions and the cultural background, as well as whether a 'trusted' institution is making the case.

One particular area of dissonance is about the perceived insecurity of research facilities. Much academic literature focuses on the possibility of malicious intruders seeking to damage data (see below), and this is also reflected in media debates. In contrast, the literature on managing data facilities is very clear that the evidence supports the idea of research use of data as very low risk.

More generally, the full social costs and benefits are often misunderstood outside the research community. For example, maintaining an (unconsented) data linkage spine in Western Australia led to a substantial fall in the number of research projects requesting access to identified data.

Risk aversion amongst data collectors

As with the public and media, data depositors tend to approach risk more conservatively than do the research community; this reflects the potential gains to each from research, and the potential losses to each from a breach of confidentiality. Some of these differences arise from differences in knowledge, as described above, but there are also arguments that the culture of data collecting organisations is more risk-averse.

In public health both parties are much more aware of the value of research, but this does not mean that interests are aligned. For example, a GP might see his or her primary responsibility as protecting patient privacy, rather than supporting public health at some risk to privacy, however small.

Academic perspectives on confidentiality

Almost all of the academic studies into the disclosure risk associated with the release of data use 'intruder' scenarios: a statistical expert with malicious intent attacking statistical outputs or databases to uncover confidential information. This has some value for discussing alternative protection techniques from a common 'worst case' scenario, but it has no empirical support. Unfortunately, the intruder model is popular with data owners, because it provides protection for these 'worst case scenarios'. Such models do not seek to balance public benefit against confidentiality protection, and so encourage over-protection of data.

Disciplinary differences

Data linkage can provide a spur to cross-disciplinary working because the ability to exploit data from different disciplines could encourage collaboration. However, there is the question of how to kick off such collaboration: does data linking put off cross-discipline collaboration?

3.3.4 Challenges for data linking: summary

From an operational perspective, linked data suffers many of the same problems as single-source data: on many topics, linked data is often more complicated in degree than in principle. There are some complications from linking, mostly to do with co-ordinating multiple organisations.

In terms of statistical theory, the main issues of data linking have been solved, and the main remaining problem seems the potential selection effect in the linked dataset. Practical problems such as cleaning data generally are seen as problems to be dealt with in user guides.

There are still large unanswered questions in the institutional framework. There are unresolved legal and ethical controversies, and a study of cultural factors shows that there are significant differences in perceptions between groups. Research suggests that citizens are reasonably comfortable with research carried out by trusted institutions; but those institutions themselves are not necessarily comfortable with releasing data.

Finally, the summary of literature above is dominated by the news and research from high-income countries. These findings do not necessarily translate to low- and middle-income countries, where, for example, one would expect data quality to be a more significant problem. Part of the aim of this project was to identify whether there were lessons that could be transferred between countries with different cultures, economics and models of governance. These are considered in the next part.

Part 2: Findings

4. Responses from interviews and case studies

4.1 Introduction

This section reviews the findings from the interviews, case studies, and participation in the Computers, Privacy and Data Protection 2015 conference (CPDP 2015), as well as drawing on the team's own experience in public health and data access, and informal telephone or face-to-face interviews with relevant experts. The views of these individuals influenced the report in many ways, and have sometimes been referenced directly; however, because each is a specialist in his or her area, and so easily identifiable, no comments in the report are sourced to individuals (except for the case studies). This was to done to allow individuals to speak freely, which they did. This report is based on the authors' interpretations of the opinions of interviewees, and no particular opinion should be ascribed to any individual or organisation.

In the discussion below, 'respondent' means any person who contributed to the discussion, including members of the authoring team. 'Researcher' and 'non-researcher' mean (respectively) a person who is or is not actively involved in statistical research using sensitive data.

Findings are summarised in four sections:

- Conceptual concerns: these relate to broad questions of whether data sharing and linking should take place at all; and if so, what should govern procedures: is there a social contract, should consent be the only gateway, can we determine general principles for data sharing?
- Contextual concerns: these consider the environment within which data sharing takes place: how does the relationship with ethics committees or data providers matter, what are the legal requirements, how is the relationship between data provider and linker managed?
- Practical concerns: these cover the specifics of data linkage: what do we know about effective data linking, how can data security be managed, what are the resource requirements of running a linked data service, and so on.
- The way forward: interviewees were asked to suggest ways to improve the prospects for data linking in public health research.

4.2 Conceptual concerns

4.2.1 Informed consent

There was a clear consensus that a sole reliance on narrow-informed consent (NIC) was not consistent with enabling high quality epidemiology and public health research.

Where data is collected for research purposes, 'broad consent' (that is, allowing one's data to be used for research but without agreeing in advance exactly what that research would be) was seen as an ethically-sound basis for research . Respondents were confident that appropriate safeguards could be put in place, and data subjects were seen to be able to make sensible informed decisions. Broad consent could also be helpful in getting public support for research, and vice-versa. For example, in Sweden the extensive public support for research use of data allowed broad consent and an opt-out to be seen as the norm (see Annex C).

Much public health and epidemiology research also uses data collected for other purposes, for example from administrative procedures or statistical data collections. For research uses of these data, there was a very strong consensus that a requirement for explicit consent would be statistically extremely damaging for epidemiological studies; and that arguments for self-interest and solidarity provided a robust ethical basis for public bodies to allow research use of data without explicit consent. There was a widespread perception that proponents of explicit consent as the only ethically valid position are typically unaware of the statistical implications, the practical problems, and the potential for the process of gaining consent itself to be privacy-endangering.

Respondents were keen to point out the practical difficulties in collecting informed consent from, for example, Census data contributors, or for designers of cohort studies to detail the projects that their data might be used for. It was also recognised that, when administrative data was being collected (for example, registering cancer treatment), the focus of the practitioner was on the administrative task, not the research potential.

It is emphasised that where existing datasets have been collected under consent (whether narrow or broad), the terms of that consent must always be respected. In addition, no participants advocated ignoring NIC completely; but the dominant view was that NIC is 'good to have', rather than necessary, and that due care must be paid to the statistical impact of gaining consent. Public health research, and linking data in particular, would not be workable without a way to provide access to some data sources without the need for consent. Researchers were keen to give examples of where public health research would not have been feasible if consent was the only gateway.

It was also noted that 'consent' is often used as it is a simple yes/no question, but this is not the case. For example, is consent genuine when a power relationship is brought to bear? GPs are in a strong position to influence the views of the subject; is giving or withdrawing consent truly a reflection of the free will of that patient? One researcher described a legal proposal in Belgium (subsequently overturned) that even employment contracts be treated as non-consensual because of the imbalance of power between employer and employee[3]. Researchers are generally more in favour of opt-out schemes rather than opt-in: as people tend to accept the default option, opt-out reduces the chance of statistical bias. However, some researchers from the UK noted that idea of 'opt-out' had been contaminated by the recent public relations failure of care.data. One case of particular interest is Sweden, which generally operates via opt-outs but uses in opt-ins for cases with very small or particular populations (see the case study).

4.2.2 EU legislation

Almost all the interviewees working in Europe cited concerns over the European Parliament's proposals in relation data protection. This would require consent for all health research except

> *"research that serves an exceptionally high public interest, if that research cannot possibly be carried out otherwise"*[4]

[3] The draft EU regulation also proposes that an employer cannot use employee data for non-employment purposes via consent, as the disproportionate power relationship means that the employees are not deemed truly free to consent

[4] Committee on Civil Liberties, Justice and Home Affairs, European Parliament (2014) DRAFT REPORT on the proposal… Amendment 328 relating to Article 81 paragraph 2, p195 (emphasis added) http://www.europarl.europa.eu/meetdocs/2009_2014/documents/libe/pr/922/922387/922387en.pdf

This was seen as disastrous for epidemiology in Europe, effectively ending observational analysis. First, the 'exceptionally high', while undefined at present, implies that the spirit of the law is for non-consensual research use of data to be very unusual, compared with widespread use in current epidemiological studies. Second, the phrasing 'cannot possibly…' provides no space to argue for a better solution, in comparison to current data protection laws which typically specify behaviour that is 'reasonable'. Member states in theory can still create a local research exemption, but it is thought this will be difficult to use.

Some respondents also noted the impact of other EU judicial decisions on privacy, particularly the 'right to be forgotten'. There was uncertainty about the impact of such decisions (at the moment these seem to relate only to search engines and other information distributors, and not the source data), but there was concern that an atmosphere might be developing which will make long-term analysis much harder.

4.2.3 Whose data are they?

Several respondents noted, in different ways, that there are two perspectives on data access:

- Default-closed: Don't release any data unless it can be shown to be safe and lawful;
- Default-open: Release data unless it can't be done safely and lawfully.

Whilst seeming to say the same thing, these reflect two very different perspectives on whether data should be made available for research (and hence how easy it is to get access to data). When asked, most researchers identify default-open as the preferred case, but default-closed as the case they experience in practice. A notable exception is the Nordic countries, where default-open seems to be accepted in practice as well as in theory (although some think there has been some reverse in recent years). The UK is generally seen as closer to default-open currently, although not all agreed[5].

4.2.4 Managing the social contract

There was clear consensus that allowing a research exemption from NIC means that the rights of the individual need to be protected by publicly accountable processes. Broadly, this translates into developing a system which can balance private need against public benefit, is transparent, is fair, is accountable, and reflects the wants and needs of society. Respondents were comfortable that this could be managed, and several researchers pointed to historical evidence of good practice.

In some conversations research ethics committees were discussed as if they were representing the social contract, but this may have been shorthand. No-one suggested that an ethics committee by itself could manage the social contract; they were there to address specific cases..

Some interviewees noted that the public also seem relatively relaxed about letting private businesses use data for research, as long as there is a substantial public benefit in the end. This doesn't reflect published research which generally shows the public more suspicious of private data; but, as noted in the literature survey, these answers tend to be sensitive to the way questions are asked.

[5] It was reported to the team that Denmark has moved formally to a default-open, broad-consent, opt-out framework for all public data; we did not have time to investigate further the drivers for this, or the impact on public perceptions.

4.2.5 Country differences

Much of the conceptual discussion was focused on Europe, as the forthcoming data protection regulation appears to be at the front of many European respondents' minds. One area of discussion is that, although there is the common European Data Protection Directive (1995), countries choose to implement it in many different ways depending on their national culture. For example, the current research exemption appears to be more important in the UK than in Germany where broad consent is the preferred route. Countries also are generally reluctant to share or link data abroad, despite the presence of EU-wide legislation and a relatively homogeneous attitude to privacy.

Some respondents found this frustrating: the potential for international research promised by harmonised regulation does not seem to be realised, or at least not at a desirable level. From a scientific perspective, international data sharing has significant benefits, including the ability to identify cultural factors in public health, and increasing observations for rare diseases. A common approach to data collection is therefore seen as desirable.

The more common view (from non-European and European respondents) was that the ethics of research data sharing and linking need to reflect the interpretation of the social contract in specific countries, and so may legitimately differ. One respondent suggested that German attitudes to data access are a response to experiences under totalitarian governments in east and west. In the US and Australia, data sharing procedures are constrained by the political consensus on the balance of power between states and federal government. In the LMICs, the need to establish a firm footing for data linking and sharing in a particular country means that international comparability is very low on the agenda; it could even be counter-productive, by disrupting a system designed for a country's specific culture.

4.2.6 Privacy and risk

Almost all respondents expressing a view on the 'trustworthiness' of researchers had little time for the theoretical literature treating researchers as 'intruders'. It was noted that the worst breaches of data security came from the day-to-day operations of the data collectors (poor IT procedures, staff selling stories to newspapers et cetera). In contrast, there was no evidence of researchers misusing data maliciously. Questions about whether researchers could be 'trusted' with the data therefore should be answered "yes", with reference to the long history of public health research.

There was more concern over the growing number of privacy breaches from hacking into organisations – even if the researchers are trustworthy, how good is their data security? Should data be stored in the cloud? Some participants expressed their doubts about systems being protected against hackers of today rather than those of the future. For example, current encryption methods are largely unbreakable, but will that be the case in the future? One interviewee rejected the question, and argued that the bigger picture should be seen – computers could protect privacy far better than paper records; the latter are much easier to steal and much more likely to be identified, whereas computer records are more likely to be pseudonymised, especially for research use.

A few respondents noted the specific concern of sharing DNA samples: these uniquely identify an individual, and also close relatives. However, unlike other identifiers such as name and address, your DNA profile is permanent and unique. If an unauthorised person gets access to that information, it is

not possible for you to change it. This would suggest putting DNA samples in a higher class of risk than other health data and demographic data.

Those who understand the IT issues seem to worry more about risk – the more you know, the more insecure IT systems seem to be. In contrast, health researchers seem to worry more about utility – the more you know, the more you focus on statistical concerns.

Pseudonymisation addresses some privacy risks – it lowers the likelihood of breach by mistakes by researchers or in IT systems. However, some researchers were concerned about the quality of pseudonymised data; in particular, if linking takes place on pseudonymised data rather than identified data, the match quality is likely to be lower.

Overall, interviewees noted that sharing and linking data does produce some privacy concerns; however, there has to be a balance of risks. Research use is acknowledged as potentially risky in theory, but demonstrably low risk in practice. Not releasing data does remove the risk of those data being misused, but it also increases the risk to society of public health not being based on evidence. To paraphrase one conference presenter: windows let burglars in, but who would live in a house without windows?

4.2.7 Principles of data access

A number of interviewees noted that some bodies were going down the route of 'principles-based' accreditation. For example, in the US there seemed to be a focus on approving the aims and methods of data linking projects, and then regulating whether one met those methods, rather than specifying the methods in advance. Similarly, in the UK, there are a number of 'principles-based' concepts circulating which have been used to define systems as well as methods and procedures. Several countries seem to be in the process of defining 'principles of data access', most of which seem to start from the default-open, social contract perspective (for example, a proposed system in Australia).

One interviewee noted that the more sophisticated access models seemed to be in countries which have more openness to sharing – particularly the Nordic countries and the UK. It was suggested that this is not a coincidence – the more open and flexible you want your approach to be, the more effort you have to put into justifying it. However, it could also be argued that this is in line with the move to a more principles-based approach, which encourages both flexibility and clarity of purpose.

One advantage of the principles-based approach is that it focuses on the aim of the system. When considering getting agreement from multiple organisations (or countries), agreeing on the aims and setting standards to meet these aims can be easier than trying to agree on specific technologies or rules. However, while there was relatively widespread agreement that a 'principles-based' approach could pay dividends, there was no common typology. 'Principles-based' was most often a description of a strategy applied by a particular body, rather than a general approach.

4.2.8 Conceptual issues – summary

The key concern on conceptual issues was the issue of whether narrow informed consent should be the primary basis for research, or whether a research exemption is at least as important. The

overwhelming consensus was for the latter; as such the forthcoming EU regulation was causing great concern amongst the European interviewees.

The need to change the tone of the debate is clear from the discussion about default-open versus default closed, and the apparent preference for principles-based planning. Both of these seek to put the objectives of data at the forefront of decision-making, on the understanding that good practice would follow (as opposed to trying to set up an ethics committee without reference to the social contract, for example).

These two issues were often seen as part of a more general problem: the failure to use evidence in decision making, particular in respect of the long and successful history of research access to data in a variety of situations.

Finally, some interviewees expressed concern that country differences were not being legislated away; however, most respondents saw that as a fair price for ensuring that the social contract reflects the situation of particular countries.

4.3 Contextual concerns

4.3.1 Public attitude and trust in institutions

Both actual and perceived public attitudes are important to successful data sharing and linkage. Some interviewees argued that data depositors are often concerned about a perceived 'public backlash' against linking data. More importantly, it is not clear that this is seen as something that can be changed. A number of interviewees seemed to accept that this is the way it is; data owners, lawyers, GPs, governments are all risk averse and you have to work with them.

However, several respondents cited surveys of the public, and occasionally their own experience, to show that this assumption is not justified. It is possible to change attitudes, albeit through a slow process. Public concerns about data access are recognised as being very specific to the question asked; education is seen as very effective, and there is a general belief that the public is much more relaxed about data access than they are portrayed. In the specific case of health data, the increased sensitivity of the data is balanced by the fact that health organisations are generally seen as the most trustworthy.

In LMICs, the reputation of the organisations seemed to be even more important, and a great deal of time was spent in building the relationship with the local community; this was an essential investment in any data linkage project (see, for example the case study on the ALPHA Network).

There was a concern among respondents that media reports affected the way in which organisations were viewed. Coverage of data linkage in the media is almost always focused on potential problems; interviewees were much more likely to reference media reports that showed flaws in data security or use, rather than positive media reports about the value of data linking. It was recognised that this was, to an extent, a natural function of the media (good news is rarely interesting), but there was frustration that this seems undo the public's natural willingness to trust health organisations.

4.3.2 Relationships with data depositors

A key element of success in data sharing or linking that comes to the fore throughout the case studies is the importance of building relationships with data depositors. If the data depositors and research data manager share a common goal, getting agreement on how this is done is greatly simplified. Some interviewees noted that agreements to share data at a high level may not translate into practical co-operation with the GP who has to hand over patient data, for example. A good relationship with the depositing organisation can help to make sure such problems don't occur, or are sorted out quickly when they do.

Good relationships can help to build a reputation as a 'trustworthy' institution. Perhaps most importantly, data depositors and research data managers working together can help to defend data linkage against challenges to the ethical basis of the research. In contrast, a data depositor who only has a lukewarm relationship with a research data manager may not be willing to expend much energy countering privacy advocates.

A further key message was that personalities matter, but in different ways in different countries. In HICs, building a good relationship with a person was important, but less so than having good relationships with the organisations in general. A concern which was reported a number of times was that the personnel in the data depositors changed jobs frequently; a new person coming into the job might want to change things, and so any personal relationship was of limited value. The best way to achieve long-term stability was to make sure that processes were agreed.

In contrast, in LMICs personal contacts appeared to be more important. This may reflect the lack of processes for what may be, in many countries, a novel approach to the use and management of data. It may also reflect that personal authority carries more weight in some societies, and so the approval of senior figures is essential to any data access.

When linking data, there was a feeling that good relationships with data depositors could be very helpful in bringing more data in. For example, if some data depositors are comfortable with the way an organisation is handling its data, this depositor can be asked to help persuade other data depositors. This does seem to be successful, particularly when dealing with government departments which put a strong value on precedent. However, this can work the other way, with one observation of an 'arms race' amongst government departments (each trying to prove that its data was more sensitive than anyone else's).

There was a feeling across all researchers that sometimes data depositors do not see the value of their work, even amongst the medical community which is seen as more supportive of research. Some researchers remarked that data collectors were concerned about 'sensational research' which might bring research use of data into disrepute, dispute a lack of evidence for this. In addition, several researchers working in LMICs noted that the lack of demonstrable local outputs from linked data research made it hard to motivate high-level support for the benefits of research. This was suggested as a potential factor in the lack of linking in Bangladesh, for example.

4.3.3 Ethics committees

In the social contract model of health research, research ethics committees (RECs) are an essential element. They can provide evidence that the public interest is being guarded appropriately. In the

consensual model of public health research, the role of the REC changes to focus on whether the research is good or not, rather than whether it should be done at all. Respondents with direct experience of presenting cases to ethics committees noted that RECs spend relatively little time on the legal aspects of data access or linkage – if the project was clearly unlawful it wouldn't have been proposed, and so the public value of the research is a far more important topic.

The relationship between researchers and RECs was difficult. Some saw them as a significant block on research, needlessly delaying work and ticking boxes rather than actively assessing the quality of research. Phrases such as requiring only 'high quality research' were cited as particularly irritating: given the uncertainty of research, how would that quality be assessed at the application stage? And who would submit a proposal for 'low quality research'?

Several researchers argued that RECs are too focused on the 'costs' side of the social contract, and not enough on the 'benefits' side – that is, they are not filling their role in the system. However, others argued that RECs see their primary role as the protection of individuals from harm in specific cases, rather than balancing costs and benefits in society in general.

Generally, however, RECs were seen as a necessary part of the whole public research framework; and there was an awareness amongst researchers that what they regarded as fussiness or risk-aversion could be seen as appropriate due diligence on the part of the approvals board. Complaints about REC seemed to be about implementation rather than the underlying principles.

What did come out strongly from the case studies were the situations where ethical approval had become tightly integrated into the whole system for data linkage, largely by identifying at the design stage the goals of the project, how the project was going to handle the data, and so on. While not quite a rubber stamp, having agreed at the beginning what the purpose of the project was, it was relatively straightforward to set up quick but robust ethical approval processes. Most importantly, the successful operations also vested authority in the ethics board to approve projects without further reference to data depositors or other bodies. In some cases this was because delegated authority had been agreed, in others this was because data depositors were represented in the process in some way. The most successful procedures also resolved the problems of competing jurisdictions; for example, by taking the approval of another university REC as evidence which could be accepted at face value, and did not need to be re-presented and re-evaluated. See, for example, the case study on SAIL (Annex C).

This integration of the ethics approval process seems to work best on dedicated facilities such as archives, where the types of projects are similar and frequent. In the most successful cases, ethical approval was fully integrated into the system, rather than being something external. This meant it was possible to build relationships with the REC members, another factor in the successful models.

Finally, researchers with the most positive relationship with RECs, highlighted the need to 'educate' the REC. This was particularly the case if the ethics approval was external to the project, as it was recognised that the REC would not necessarily understand the nuances of the research or the value of it.

4.3.4 The risks of research: 'insider' versus 'outsider' perspectives

Related to all the above three problems is the issue of how the riskiness of the research access to data is perceived. When considering the safety of data access solutions, almost all respondents noted that data security was managed very effectively in the systems they had access to. This is perhaps unsurprising, but many also cited the general low-risk nature of research data access. There was a strong sense that there were risks in theory, but that there was ample (and demonstrable) experience to manage research data access safely and sensibly.

At the same time, several respondents expressed annoyance or bafflement that others did not understand this: that risks of access kept being raised by data depositors, RECs, legislators and others, despite the absence of any supporting evidence. However, within the data community there was little discussion of the topic; this characterisation of the situation was treated as common knowledge.

This raises the possibility of an 'insider-outsider' split in perceptions. Research data professionals see no need to discuss the evidence base for safe use, as it is common and widespread knowledge. Meanwhile, those who do not consider issues of data access on a regular basis have no need to familiarise themselves with practical details.

4.3.5 Working with government

Much of the data used in epidemiological studies comes from government sources. Experiences of this differed between countries.

In HICs, three main problems were identified. Two have been discussed already: the high turnover of staff (and hence limit on ability to build relationships), and the potential for turf wars between departments anxious to maintain their authority.

The third issue was the variability of attitudes to data linkage. It was noted by many respondents that organisations working in public health tend to have a much more positive view of research and a willingness to support projects. In contrast, government departments dealing with socio-economic data appear to be much more cautious about allowing their data to be linked, despite the fact that health data is likely to be much more sensitive. One reason for this difference may be that health research can have a clear simple payback (change of medical procedures or dietary advice, say), whereas socio-economic research often has, at best, a very diffuse impact (learning more about the unemployed doesn't directly lead to lower unemployment). A second reason may be that senior health care professionals are likely to have practical experience of research projects, unlike economics research, for example, which is more typically outsourced to external academics.

In LMICs, two different questions arise. The first is the relative importance of health research in relation to other government priorities. Typically there are other pressing needs. If the country has had little experience of epidemiological research, then it may be hard to demonstrate relevant and useful impacts. Hence, it can be seen as a low priority. In addition, such research might be expensive and dependent upon external support (financial and technical). Finally, if personal authority is important and it is unlikely that ministers will have relevant experience, there may be no chance to find a 'champion' at the top.

A second problem relates to the perception of the government itself. Health organisations needing to link data and working with an unpopular or untrusted government may find themselves tainted by association. It is not clear how this has affected research, but organisations working in LMICs see this as something to be aware of and concerned by.

4.3.6 Researcher attitudes

Finally, it was noted that researchers have an important role to play when developing a positive research environment. A number of commentators suggested that researchers can be over-protective of their data, discouraging sharing, and limiting linking.

Interviewees recognised that this was a natural reaction – research data managers have typically spent a great deal of time collecting and combining data, and want to exploit full value from these data. In HICs in particular, academic publication of research findings is often seen as a premium output, whereas articles about data collection are of limited value. Hence, there is a strong incentive to maintain control over one's data; it was suggested that some researchers do not trust others to give them appropriate credit for data collection, or that the researchers may be afraid that others will seek to find errors in the data. One respondent simply cited 'professional jealousy'.

It could also be argued that this is an efficient way of working – researchers collaborating with the research data managers can exploit the latter's knowledge of the data – and may be necessary to get access at all: for example, in Sweden foreign researchers need a co-researcher based in Sweden to satisfy legal requirements. However, respondents noted several cases where the requirement to work with the data team seemed less than co-operative; and the team is aware of at least one case where the data manager insisted on research groups having exclusive access to the data (with, of course, the data manager's research group prioritised).

This phenomenon was not limited to high-income countries, with respondents noting similar problems in LMICs. This may seem surprising: the lack of data for analysis puts researchers in a strong position to ask for collaborative work (as opposed to just handing over data); a lack of technical skills may also encourage joint production of analysis. This precisely the problem that the INDEPTH Network (see case study in Annex C) was designed to address.

Some commentators noted that this might be the case within projects, but was less likely to happen between projects. As many of these projects are externally funded, it might be that scope for sharing is strongly determined by the attitude of the funder, rather than the individual researcher. However, there was more of a sense that researcher 'protectiveness' dominated.

4.3.7 Contextual issues – summary

Key to the successful operation of any linked data project is the relationship with others: the public, the researchers, the data depositors, the RECs. The experience of those consulted supports research findings that the public generally are very supportive of medical research, something that may not be acknowledged enough. That support is closely related to the trust in the institutions, and medical organisations tend to be well trusted.

Relationships with data depositors and ethics committees can make the difference between a successful project and a failure. For HICs, strong organisational links seem to make the difference with data depositors, whereas for LMICs personal links seem to matter more.

There was a widespread recognition that researchers can be part of the problem too – not everyone is as free with the data they hold as one would wish. This is understandable, but is perhaps something that funders are best placed to tackle.

4.4 Practical concerns

4.4.1 Data quality

In high-income countries, data quality was discussed and specifically addressed in the interview schedule, but it came relatively low on the list of concerns, and interviewees focused more on institutional issues. Data problems were identified, and some were absorbing a lot of time; but these were generally seen as practical matters, not major barriers. For example, one researcher gave the example of a 'smoker' being defined 'hundreds of ways' in the different code systems they had to use. This was mildly annoying, and consuming the researcher's time, and yet it was not of the same order of concern as, say, ethical approval. Similarly, researchers in HICs were more likely to comment on the quality and compatibility of metadata, not whether it existed at all.

For interviewees involved in LMICs, data quality was a much higher concern – it was noted that:

- the identifying variables may not exist at all: link fields such as name and address may be missing;
- the identifying variables may be misreported, perhaps deliberately: for example refugees, or for patients presenting with socially stigmatising illnesses;
- the variables of interest may be missing or of low quality - notable problems arise where the subject is illiterate, or is not able to communicate in his or her first language;
- the lack of an integrated health-care system (or other systems) may mean that data can be specified in many different ways ;
- a lack of triangulating data sources may make it difficult to validate any link; for example in HICs, studies show how linked data can highlight the gaps in administrative data, but in LMICs the main source of data appears to be from specific collection, and so there is limited opportunity to evaluate data sources.

These problems are not unique to LMICs, but they seem to be much higher on the respondents' list of perceived problems. There may also be an element of a vicious circle: without demonstration of the value of linked data, the systems to achieve it may not be seen as a priority. In Bangladesh, for example (see the case studies in Annex C), it is questionable whether the lack of data discourages linking or the lack of demand for the research discourages the effort to create linkable data.

Some projects which have deliberately set out to collect and then link data from multiple sources have been effective in enforcing common standards; in general, however, the key concern was that data quality needs to be improved substantially, both to allow links to be made, and to do useful analysis with that link.

Most interviewees also highlighted South Africa as an exception, particularly within sub-Saharan Africa. Although there are still substantial data problems, systems have been set up to create and collect unique identifiers, and there was more confidence in the underlying data; there was also some confidence in where biases in data collection were likely to come from. This may also explain why institutional (rather than statistical) issues were more likely to come up when discussing data linkage in South Africa compared to other LMICs.

4.4.2 Timing and funding

Several interviewees commented on the time to get projects approved, which ranged from months to years. One researcher claimed that approval for his project, for which the data and link pre-existed, took 40% of the entire project time. Similarly, researchers noted that data cleaning, linking and preparing for use took much more time than non-researchers expected.

Because the various processes have to happen in a particular order, delays can have significant knock-on effects. One study reported that negotiations for data access took almost four years, necessitating extensions to grants and the need to find productive activity for those who were expected to work on the linking. It may be difficult to judge when it becomes appropriate to begin appointing data linking staff and commissioning IT systems: too early and the resource sits idle, too late and the project is further delayed.

Whilst all research projects have the potential to slip their timetable, linking data clearly provides increased risk of cost over-runs. It is difficult to assess the true cost of such additional expenditure, as the costs may be hidden. For example, in the four-year delayed project noted above, it seems likely that alternative activity was found for those employed in expectation of access being given. In contrast, the project manager spent far more time than planned in negotiations, and would have been unavailable for other work.

The interviewees had not seen any systematic analysis or meta-analysis of cost overruns associated with linking projects, and so the impression that linking projects are highly vulnerable to uncertainty and prone to cost overruns is based on anecdotal evidence. Moreover, it is not possible to say that such costs are excessive. Researchers or data linkers may bemoan delays to access; but data owners would be failing in their duties if they did not carry out necessary checks because they were under pressure to meet an external timetable. From the data owners' perspective, higher costs may be a reflection of due diligence in the face of greater uncertainty or sensitivity.

However, this has an impact on funding. Many interviewees thought that funding streams were not well suited to data linking projects, as the time to create data was always underestimated. Funding was also often focused on the outcome of the project, not the data capture. This mean that much of the investment in data gathering for a project could be lost as the funding finished and the researcher moved on.

For some interviewees, delays in getting data made them very wary of involving PhD students – some would only advise a student to work on a linked dataset if the link was already complete, proven and approved for research use. Some were more relaxed, but these interviewees mostly had, or knew of, proven systems which could deliver linked datasets in a timely manner.

Many interviewees felt that the problem was to separate out funding for data creation and linkage from analysis; this would mean appropriate costs could be allocated (for example for software tools or metadata development), and allow for project plans which focused solely on data outputs to be drawn up. This would also allow outputs to include data publications, rather than academic research articles. Most importantly, this would mean that funding data creation could be seen as an investment (and treated as capital expenditure), rather than an expense (current expenditure) as it is now.

It is worth noting that most of the comments about delays in linkage projects came from interviewees working in high-income countries – this seemed to be of lower importance in LMICs, again perhaps because this is not the main problem.

4.4.3 Capacity-building

All interviewees recognised the need to build capacity, although there were some slight differences between countries. The key issue is training linkers to link data, and users to use that data effectively: dealing with multiple streams of events in longitudinal data was given as an example of a specialist skill which could only be developed by working with this type of data.

One problem highlighted is that much of this capacity-building comes through experience, which requires a long-term commitment. Because data creation does not generally lead to high status publications, becoming a data specialist may not be an attractive option for junior researchers or those keen on developing their publication record.

It was noted that for LMICs much of the expertise (and so training) in data collection, management and linkage was coming from HICs. This was identified as a way of building up good long-term relationships, not just the direct training effect (see, for example, the ALPHA Network case study in Annex C). Some HIC organisations were also using remote technology to develop the experience of their LMIC partners.

4.4.4 Storage models

While HIC funding agencies often require data deposition in an archive as a condition of funding, this may not be feasible for identifiable linked datasets.

There were a wide variety of models used for holding data for linking and analysis. Some projects tended to keep data separate and only link for the specific project. Others promoted data archives to store complete datasets for re-use. Another model was to keep the link fields permanently, but only pull in the data of interest when necessary

The models have different advantages. For the data archive, getting ethical approval for the archive in the first place is key; future requests for data should then be simplified.

The link-as-necessary model has the conceptual advantages that data is not linked unless it needs to be; the downside is that this can lead to much slower ethical approval.

All interviewees agreed that maintaining a master link key was essential to making the data extraction process quick and reliable. This allowed linking errors to be fixed on a cumulative basis, avoiding repeated errors (see, for example, the WADLS case study).

Interviewees noted that there are multiple models of managing data in use. There is a need to recognise that different data, for different purposes, can be managed in different ways, even in the same country. For example, the UK has a range of different facilities for analysing linked data, at various levels of detail. There are a number of conceptual models being used to define such relationships, such as 'zone models' or the 'five safes' framework. Some interviewees mentioned 'safe havens' (also called data enclaves or research data centres) as a nice idea, but felt there was too much variation in what constituted a 'safe haven' for it to be a useful definition.

Some facilities were also beginning to make use of remote access technologies, although this is still comparatively rare for these sensitive linked datasets. However, few saw themselves as being seriously constrained by restricting data access to a fixed facility; for example, the Scottish Longitudinal Study (see case study) had experimented with a range of tools to make the occasional visits to the restricted-access facility more productive.

Finally, it was noted by some that the permanence of data stores has not been resolved. As data accumulates, should any of it be unavailable for research? Much progress in data linking has come in the last ten years or so, in line with developments in computing. What are the ethical implications of cloud computing, for example?

4.4.5 Practical issues – summary

Data quality is a major issue for LMICs, whereas for HICs other practical problems seem to be more pressing. The opposite seems to be the case for concerns over timing and wasted resources. The case of South Africa suggests that there is a natural progression from operational problems to more statistical ones as data linking increases and becomes more the norm. Given the longer experience of HICs in data linking and managing, there are gains to be made from sharing information about skills, data facilities storage models.

4.5 Ways forward

When asked to suggest ways to improve or benefit more from data linkage, most responses came down to more money – often, money specifically targeted at data management rather than on producing research outcomes.

However, there were also some suggestions that a better understanding of each other's roles (clinician, researchers, data managers, ethicists) would allow for more realistic expectations of what could be achieved for data linkage, and over what time scale.

Part 3 Conclusion and recommendations

5. Summary of findings

5.1 Broad conclusions

The broad conclusions of the report can be summarised as follows:

- Theoretical or statistical challenges for data linkage can generally be seen as solved, at least for practical purposes
- Practical issues still exist, and are much more important in LMICs where data quality is lower
    - Good consistent identifiers substantially improve outcomes, but should not be pursued at the expense of the variables of interest
- There is a need to ensure that decisions about linkage are well-informed and evidence-based
    - Narrow informed consent alone is not a basis for good epidemiological research; some form of workable research exemption is necessary
    - There is ample evidence to show that the social contract can be managed effectively
    - There are substantial differences in the ethical positions taken by those in authority, which seem more to do with cultural or institutional factors than genuine ethical matters; this variation in practice (even within countries) has a substantial negative effect on research
- The general public (at least in HICs) is very supportive of using linked data for research
    - Trust in institutions is one of the most important factors for public acceptability of research use of data, at all levels of decision making
    - Trust is fragile, but memories are short: one incident can set research data access back a long way, but only if recalled
    - The framing of questions is crucial to issues of public acceptability
- The data management community largely sees as a stylised fact that research use of data is relatively low risk, and can be (and has been) managed safely and effectively
    - For this community, safe management of data is a practical matter of designing systems, procedures and training
    - This view and the evidence base behind it does not seem to be communicated well outside that community, who are more likely to focus on theoretical risks
- Cultural issues are important in determining the success of a project:
    - Personal relationships and personal authority can go a long way to resolving (or creating) problems.
    - Turf wars and power relationships can create reasons for excessive regulation.
    - Some academics are resistant to sharing data, even where funders require it – there is a desire to exploit one's monopoly
    - This was identified as a more significant barrier in HICs, compared to LMICs
- Incentives to manage and link data are weak
    - There are few incentives to specialise or develop expertise in data, per se
    - Transferring knowledge to LMICs is a resource-intensive process
    - Data linking is a long process which should be better viewed as an investment in a cumulative store of knowledge

5.2 Response to initial questions

The project tender identified twelve questions. We use them to consider what conclusions can be drawn from the analysis and interviews.

5.2.1 General queries

What are the potential benefits (including impact) and opportunities of linking research data?

- The public health impact of linking data is demonstrable. A very small selection of the many useful outcomes, illustrating different aspects of data linking, were presented in part 1 of this report. It is clear that not allowing data to be linked would severely impede public health research.
- The benefits differ between countries. In HICs, benefit comes from the enormous range of data held in medical and non-medical data sources. In LMICs at present the benefit seems to come from building up sufficiently large samples to carry out effective analysis, and potentially to identify country-by-country differences.

What are major challenges (e.g. technical, ethical, legal, financial, cultural) that prevent these benefits being realised and how might those challenges be addressed?

- The statistical barriers can be considered largely solved. While there is continuing interest in the area among statisticians, for practical purposes there are no unresolved problems.
- Data quality, of both link fields and variables of interest, remains a significant problem in LMICs – perhaps the most significant problem. Whilst it is also a problem in HICs, generally it is comparatively minor.
- Ethical and legal barriers to data linkage generate lively debate in HICs, but there is likely sufficient common ground upon which to resolve them. An exception is the proposed new EU data protection regulation, which is seen as potentially fatal for observational research.
- For LMICs, ethical and legal issues were not raised as significant, perhaps because these are less relevant at the moment. Most of the projects identified have got specific arrangements for their situation, which works. To paraphrase one interviewee: getting my data linked is fine; getting a general strategy for linking data is dead in the water.
- In HICs, institutional barriers seem far and away the most prevalent. The heterogeneity of solutions shows that there are many different ways of solving problems, and yet problems still exist. Some of these can be put down to cultural preferences, but in many cases it seems that either breakdown in relationships or a failure to define a vision has led to unnecessary delay or cancellation of projects.
- One recurrent problem in HICs is that decision-making bodies are often unwilling to show faith in either the decisions or systems of other institutions (for example, in taking a group approach to ethical approval); it could be argued that a lack of personal knowledge engenders a lack of trust.
- For LMICs institutional issues focused around the need to connect with the right people. However, in South Africa, institutional issues more familiar in HICs were emerging. This may be because data quality issues are now less important, or because South Africa is trying to develop a more strategic approach to data, or a combination of both. In any case, it provides an interesting example of how barriers change as the data environment evolves.

- Training people in the collection, linking and analysis of data appears to be concerning many people. At the moment, a lack of data professionals and trained researchers does not seem to be holding back projects excessively, but there is a concern in all countries that there is insufficient investment in data skills.
- Funding is thought to be problematic, partly because it typically focuses on research outcomes rather than data creation, even though the latter can take most of the project time. Funding data resources separately would also lead to better management and accounting – at present there is no idea, for example, of what delays in approval cost. It would also allow some current research expenditure to be seen as capital expenditure.

What specific challenges exist in relation to: linking, harmonising and pooling data across national boundaries?

- This was not investigated in detail, due to time constraints. However, in general it seems to reflect institutional problems. Sharing confidential data across national boundaries is almost impossible in most HICs. Data linking has no effect: as individuals are not expected to live in multiple countries, pseudonymised data can be used.
- The case studies in LMICs seemed to show that data sharing is much more feasible; this may reflect the fact that these projects have been negotiated in great detail, and are not trying to set general precedents.

How is effective data linkage defined conceptually and in practice?

- There is no conceptual definition, as it depends entirely on practice and the purposes for which the data are being used. The assumption is that no data linkage is perfect (even with common link fields), but, in line with data quality generally, is it good enough? This is a judgement call, but in HICs there is often some other data source which can be used to triangulate linking. For LMICs this may be more of a problem as the data being collected and linked may be the only data available.

What lessons can we learn (for research funders, researchers, policy makers and health practitioners) from existing data linkage initiatives in terms of the systems that they are using and the training that they are providing?

- For HICs, the most successful cases have had a strong design element where all the difficult questions about ethics, legality, and jurisdiction have been assessed in advance. The culture is more likely to be default-open, and an evidence-based approach to risk is more in evidence.
- For LMICs, building personal relationships and leveraging external contacts to piggyback on training and expertise appear to be the main lessons.

What best practice principles should be adopted and what practical solutions could be considered?

- See recommendations below

What is the relative position of different fields in relation to utilising data linkage e.g. biomedical, health, economic, environmental, social data to produce evidence to support policy and delivery of health services and medical interventions?

- Generally, the medical/public health profession is strongly supportive of data linkage, as it addresses many known problems in, for example, case control studies, as well as solving issues which can't be achieved via experimental methods. Other professions seem more concerned about sharing and linking data, possibly through lack of contact with research or because of the diffuseness of research benefits in those fields.

5.2.2 For low- and middle-income countries:

Do the challenges and benefits differ: (i) within and between LMICs and (ii) between LMIC and non LMIC countries? Are there transferrable lessons?

- The case studies and the literature reviews show that there are differences. However, the South African cases suggest that there might be a natural path through which all countries move: use personal contacts to get some data; improve the data quality; increase the scope of the project; use that project to develop other projects; develop a strategic approach; develop a country-wide strategic approach. As you move through the stages, you concentrate less on the practical problems; the institutional problems start to loom larger as you are now starting to change society more generally.

Are there specific approaches that have been effective in LMIC and non LMIC countries?

- In the LMICs, working on personal relationships to develop very specific programmes looking at specific issues has been very effective; in the HICs, programmes are increasingly looking at good practice in other fields and abroad to find arguments for more efficient and flexible processes.

5.2.3 Future trends:

What are the new and emerging data sources which have the most potential in relation to data linkage in the field of public health?

- No new data sources were identified. Whilst there was discussion about 'Big Data' and greater use of administrative data, the main perspective was that there was sufficient knowledge and experience to handle any such developments.

What are the new and emerging technologies and methods that are having an impact now or in the future on data linkage in the field of public health? What are the implications for governance?

- The biggest technological development is in remote working. At its most basic, telepresence and videoconference tools allow teams to communicate and share knowledge across organisations and distances. The ALPHA Network (see case study) uses very simple technology to carry out multi-country analyses. At the top end of the researcher experience remote research data centres (or data havens, or enclaves) allow researchers full access to data from any location (including poorly connected LMIC locations, as such applications require very little bandwidth).

- On methods, the current interest in principles-based approaches seems to have the potential to address many of the institutional problems faced in HICs; these also have more applicability to LMICs than the very specific regulations used by many HICs.
- The principles-based approach also simplifies governance.

Where might we be ten years from now in terms of data linkage?

- If all the best practices currently used in different places were applied universally, and followed the principles-based route, we would be in a strong position in HICs, and we would have a clear development path for LMICs.
- If however, fear prevails it is quite possible that HICs will step further away from data linkage, particularly if NIC becomes the norm in all but exceptional cases. This would leave both LMICs and HICs reduced to ad hoc solutions in particular cases.

6. Recommendations

Our recommendations to the Public Health Research Data Forum (PHRDF) are largely concerned with distributing useful and accurate information to change ideas about data linkage and show the possibilities to interested parties. We recognise that members of the PHRDF are major research funders, but they do not have a statutory role and they have to work within the constraints of the society within which projects are sponsored. Nevertheless, we believe that a common perspective from a critical mass of funders would substantially improve the environment for and practice of data linking.

We believe that most of the recommendations could be implemented in a relatively short period and at relatively low cost. The combined impact of the recommendations should be to change the debate from “Can we…” to “How do we…”. In countries and organisations that have made this progression, it has been observed as a slow process needing constant support and reinforcement. We would therefore like to see any attempt to address the recommendations in the short term accompanied by a longer-term strategic commitment to encourage evidence-based data planning.

6.1 Recommendations and rationale

Our recommendations to the Public Health Research Data Forum (PHRDF) are largely concerned with distributing useful and accurate information to change ideas about data linkage and show the possibilities to interested parties. We believe that a common perspective from a critical mass of funders would substantially improve the environment for and practice of data linking.

Our recommendations are grouped around two topics: setting the conceptual framework, and finding solutions to practical problems.

6.1.1 Set the conceptual framework to control the debate

The aim of this set of recommendations is to change the general language of debate to make it more supportive of data linking, and provide the conceptual basis for strategic thinking on improved data access.

- ***Change the language used when discussing data access from default-closed to default-open***

The initial perspective affects where you end up. Changing the default assumption to "data should be available for research, unless there is a reason why not", for example in publications and funding calls, can change perspectives to focus on utility rather than risk.

- ***Develop and promote high-level principles for research access to data and data linking***

A number of countries and organisations are moving towards principles-based specification of security systems. At the same time, while data professionals have a good idea of what makes an effective data management system, they are often required to start discussions from first principles with data depositors or government regulators. A statement of 'best practice principles' would help to foster coherence across systems and support for research managers in specific cases. The Australian data linkage principles currently undergoing consultation[6] are an example of how these could be presented.

- ***Encourage practitioners to share their knowledge and experience of effective risk management in research access***

Data professionals see as unremarkable the idea that research access is demonstrably low risk when managed effectively, as their experience shows this to be the case. This can sometimes mean that they may not make sufficient efforts to convince 'others outside of the field, who, in the absence of experience, place more emphasis on conceptual risks and worst-case scenarios.

- ***Develop a toolkit of coherent cases, backed by evidence, which can be used for advocacy purposes in policy discussions***

Effective advocacy requires knowledge, experience and persistence, but also knowing what sorts of arguments work. Providing a 'toolkit' of resources setting out the case for data access with exemplars would help greatly those taking forward advocacy efforts around the world and help them to make a consistent case. Such resources should cover:

  - the need for a practical research exemption from narrow informed consent;
  - the high-level of public support in and trustworthiness of the research community in general and public health community in particular;
  - the risks to the public of not being able to use health data in research;
  - the safety record of research facilities.

- ***Produce guidance on best practice ethics processes which encourages collaboration and co-operation***

Again, there are some key issues of principles which it might be useful to have to hand when thinking about ethical approval – for example:

  - the difference between legal responsibility for due diligence and needing to examine all evidence oneself;
  - evidence-based assessment of risk;

[6] http://consultations.nhmrc.gov.au/files/consultations/drafts/draftprinciplesaccessingpubliclyfundeddata141209.pdf

- acknowledgement of precedent (historical and other committees)in decision-making.

6.1.2 Help resolve practical problems with specific advice on good practice which seems to work

- ***Encourage the use of remote technology to allow knowledge transfer between HICs and LMICs, particularly collaborative working tools***

There is a lot of technical skill in HICs and local knowledge in LMICs which technology could bring together. The comparative work by the ALPHA Network is one example, but it could go much further. For example, the remote-Research Data Centre (virtual safe haven, virtual data enclave) model is well established best practice in Europe and North America for dealing with confidential data; it could be adapted to allow cross-border collaborative work on less sensitive data, and the technology is cheap. The LISSY[7] system has been providing a remote job service for over twenty years with upwards of 50,000 analyses over that time, and no breaches of confidentiality. Tools such as NESSTAR[8] or other metadata systems are available off the shelf; tabulations tools are available which have built-in confidentiality protection. Finally, simple telepresence technology (web conferencing etc.) is available, in many cases for free. Along with this, protocols such as the 'five safes' model were developed specifically to allow potential data managers to consider their options consistently and in the interests of the user as well as the data depositor.

- ***Provide dedicated funding for the creation and management of data resources as a distinct element in research grants***

Funders should consider whether the data management part of projects should be identified and funded separately. This would allow data issues to be recognised and managed as problems in their own right (rather than something which holds up research), provide clearer incentives for 'data professional' to be seen as a career path, encourage post-project development by seeing this as an investment, and may allow funding to be allocated from capital as well as current expenditure.

- ***Invest in PhDs as a cost-effective long-term investment to develop data expertise in LMIC and HIC settings***

In this case we consider (1) HIC-based and supported PhDs, probably students from the LMICs, developing data linking as a specific part of their thesis, and acquiring local knowledge about LMICs targets; the idea would be that they return to LMICs on completion and take their skills with them, rather than expensive staff being sent out without good practical and local knowledge (2) PhDs (HIC and LMIC) focussing on data expertise; this directly addresses the problems of PhDs not getting involved with data because of the risk of non-completion, and building long-term capacity by making data specialism recognised and valuable.

- ***Draft guidelines for research teams on addressing practical issues in enabling data access and linkage***

This is similar to the conceptual guidelines above, but more focused on the practical matters such as:

[7] http://www.lisdatacenter.org/data-access/lissy/
[8] http://www.nesstar.com/

- what makes an ethics committee work with you rather than against you;
- why spending time developing a reputation for trustworthiness is a long-term investment;
- the pros, cons and past experience of alternative data management systems;
- effective researcher management;
- frameworks for discussing confidentiality;
- avoiding duplication of information gathering for multiple committees.

- ***Build up a record of 'useful' precedents, experience and exemplars***

Precedents have power, particularly when dealing with government departments. Again, the point is to have a toolkit of options available to support the research community. Almost every data management practice has been implemented somewhere by someone, and there's usually a 'good' example to find - the problem is that at present only data professionals may be aware of those precedents, and not appreciate their value.

6.2 Timing

It is notable that these recommendations are largely concerned with distributing useful and accurate information to change ideas about data linkage and show the possibilities to interested parties. Hence we believe that these recommendations could be implemented, at least in draft form, in a relatively short period (that is, within the year) and at relatively low cost (that is, in terms of weeks of effort rather than many months or years).

However, we also argue that there a longer-term commitment. The combined impact of the recommendations is to change the debate from "Can we…" to "How do we…" In countries and organisations that have made this progression, it has been observed as a slow process needing constant support and reinforcement until the paradigm has shifted. We would therefore like to see any attempt to address the recommendations in the short term accompanied by a longer-term strategic commitment to, for example, periodic review.

As noted above, we recognise that PHRDF members have no authority to compel changes in attitudes. However, we believe that the support of such a key group of organisations would make a substantial change to the environment for data linking.

Three recommendations are not concerned with changing attitudes, but with practical matters. In each case, it is not entirely clear what the long-term strategy should be, but there are some achievable short-term goals which may help to define that strategy.

Sharing knowledge of remote technologies is unlikely to be effective on its own; there may be a need to invest in demonstration projects. These would need to be chosen for their strategic value; however, in this project we had insufficient time to suggest cases for demonstration projects. Hence we would suggest that, as a short term goal, the information sharing is sufficient, but in the medium term identification and funding of demonstration 'remote collaboration' projects would be desirable.

Separately identifying the funding for data aspects of research proposals may run into practical barriers: for example, data creation may be capital investment, whereas research is current expenditure. Some major projects are clearly investments in data development, but for others the

boundaries are less clear. Few researchers are likely to welcome a longer application form, but a small number may support the idea where data creation is a large or risky part of the project (and hence see this as a positive development). The short-term goal for this objective would be to explore the demand for separate funding streams.

The recommendations relating to PhDs also require a more significant investment. It also raises significant questions: for a PhD in being a 'research data professional', what academic discipline should this be? Who would be a qualified supervisor and examiner? Placing the PhD within maths, statistics, operational research or epidemiology would each send a different message and perhaps have a different long term outcome.

Again, a feasible short-term target is to help identify the potential uses to which such a person could be put, and identify possible strategic collaborations which could be funded. One way would be to invite expressions of interest, and let the research community decide what it thinks is most useful.

Annex A: Overview of relevant literature

This annex provides a brief overview of the relevant literature, summarised in the main document. It covers:

- Concepts in data linking
- The value of data linking
- Problems of data linking

A1. Concepts in data linking

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. For example, a study of medical records may show that young mothers have a poor diet. However, linking this with economic information may show that the age of the mother is correlated with income and that poor diet is associated with low income rather than the age of the mother per se.

A1.1 Identifiers and identification

When linking data, variables are typically split into:

- Identifying variables (for example, name, address, medical insurance number)
- Variables of interest (age, gender, income, illness, occupation etc)

Identification means associating information with a known individual. Identifying variables can be *direct* or *indirect* identifiers. The former allow individuals to be identified exactly (for example name, or medical reference number). The latter only identify individuals in combination with other information (for example, age, gender, and occupation in combination with postcode).

Direct identifiers are typically of little interest to researchers; their value is in allowing the data to be linked. They do however allow a known individual to be associated with potentially very sensitive information. Hence good practice generally requires direct identifiers to be removed from datasets before they are made accessible to researchers.

Indirect identifiers and variables of interest often overlap; for example, age and gender can be used to identify an individual but are also typically valuable explanatory factors in any analysis. This does mean that a dataset is likely to have some characteristics which will allow the individual to be-re-identified from the data; for linked datasets this likelihood increases as the number of characteristics is increased. This can be a problem when discussing confidentiality with researchers who may not make the connection between the range of characteristics they want and the increasing identifiability of the data.

A1.2 Types of data linking

A number of techniques are available for data linking.

A1.2.1 Exact/deterministic linking

Exact (or deterministic) linking is possible where a unique identifier is shared between two data sources. For example, in the UK National Health Service (NHS) number is used to link data across NHS medical records; across organisations, the national insurance number (NINo) is used by the tax department, the social security department and the national statistics office. It is therefore possible to link information from all these sources directly. Consider the example below

Surname	First name	Address	Town	Postcode	Sex	Age	NINo
Smith	John	17 London Road	Birmingham	B1 6AS	M	42	AB264254Q
Name	Employer	Employer's address	Occupation	Ethnicity	Sex	Age	
JB Smith	Altrex ltd	Altrex House, Broadway, Wolverhampton	Surveyor	British	M	43	AB264254Q

On the assumption that the NINo is recorded correctly, only the last field is needed to link the two records together.

In theory, the obvious advantage of exact matching is that the link is certain and simple to effect. A secondary advantage is that the match field is typically a non-informative reference number. For example, the UK NINo is a random collection of letters and numbers (in contrast a UK driving licence number is unique but informative as it contains substantial information about the owner embedded in the code). This means that a non-informative match field can be circulated between research groups with less concern about identifying information being released through accidental exposure. For example, if a data set containing NINos is accidentally released, individuals could self-identify or could be identified by others using private/unlawful data sources; but compared to a dataset which contains names and addresses the risk of identification is much lower.

As well as the uniqueness of the match field, exact matching is based upon the assumption that the data are accurate. This depends upon the resources available to the match field creator, and the importance of correct matches. Credit card companies and software licences typically incorporate 'check-sums' which allow the accuracy of the card or licence number to be verified instantly. In contrast, a hospital may not have the resources to create self-checking record numbers; moreover, it may take the view that such numbers are administratively convenient but the primary check is always the name of the patient plus data of birth or first line of address.

Where the two data sources reference different points in time, there is also a requirement that references are not re-used in that period. Re-using references can cause confusion if this information is not known. For example, although NINos are supposedly unique to an individual, 'temporary' NINos are issued and re-used for some statistical purposes. These are identified with a special code, but to the unwary researcher there appears to be a surprising number of people in the UK who change sex repeatedly.

Exact matching need not be carried out on random reference numbers: for some purposes a name and date of birth might be sufficient. However, as the match fields stop being arbitrary reference

numbers and reflect real values, the assumptions of clean accurate data and unique values become less robust. One approach is to define more rules ("the first name 'John' can be represented by the initial 'J'") but this can become too complicated to be manageable. An alternative approach is to replace deterministic matching with probabilistic matching.

A1.2.2 Probabilistic matching

Probabilistic data matching is a well-established and common solution to data linkage. Name, address, age and gender (for example) are common across many data sources, whereas common reference numbers for exact matching require a degree of co-ordination between organisations.

Where a field of guaranteed unique references does not exist, or if significant errors are thought to occur in the data, *probabilistic matching* is carried out. This takes the (individually) non-unique fields and gives a probability that two records relate to the same person. Consider Table 2

	Surname	First name	Address	Town	Postcode	Sex	Age
Target:	*Smith*	*John*	*17 London Road*	*Birmingham*	*B1 6AS*	*M*	*42*
1	Smith	John Brian	17 London Road	Birmingham	B1 6AS	M	42
2	Smith	John				M	42
3					B1 6AS	M	42
4	Smith	J	17 London Road	Birmingham	B1 6AS	F	42

If the aim is to match the first record with one of the others, the match process could be reasonably confident that prospective match 1 is the same person – all fields match, with the exception that one shows an extra middle name. As this is commonly omitted from records, the match seems likely. Not putting the two together would lead to a *false negative*; that is, a true match not being recognised.

The second is less likely. All fields match, but the values are not individually unusual. It is quite feasible that at least two John Smiths, male and aged 42, exist, and so asserting that the two individuals are the same has a high probability of a *false positive*: declaring to records to refer to the same person when they do not.

The third match would seem to have a very low probability of success if only three fields match. However, in the UK a postcode typically represents 20-30 houses, and it may be a reasonable expectation that match 3 is the same as the target. This expectation would be strengthened if it was possible to check how many 42-year-old males live in that postcode.

Finally, the last seems a good match on all but one field. This could be a miscoding, but gender is relatively simple to code. Perhaps a more likely explanation is that the Smiths of London Road Birmingham are a married couple of the same age. The dissonance in this single field means that the interpretation of all the other fields needs to be reconsidered.

The score used to determine whether two records match or not is generally calculated as the ratio of two probabilities: the likelihood of a true positive, and the likelihood of a false positive. Hence a record which matches the target but which could also match many other targets might score lower than a record which does not match the target as well but is extremely unlikely to have another candidate in the data. This method is almost fifty years old now[9] but, despite criticisms of the underlying assumptions, alternatives have not proven themselves to be notably better.

[9] Fellegi I and Sunter A.(1969). "A Theory for Record Linkage". *Journal of the American Statistical Association*

Software to carry out probabilistic matching typically sorts data into 'matched', 'unmatched' and 'uncertain', with tolerances defined by the user[10]. The aim is that the person overseeing the process can focus attention on the 'uncertain' area to be confident that the matches and non-matches are valid. Clerical matching (going through the data by hand) can be focused on the 'uncertain' areas.

It should be clear that probabilistic matching is a much more subjective process than deterministic matching. It requires the person matching to take a number of decisions:

- What combinations of variables should count towards the match?
- How strict should the requirements be? Stricter requirements for a successful match reduce the chance of false positives but increase the chance of false negatives, and vice versa.
- Should inconsistent values be treated as errors?
- Should inconsistency in some variables be treated more seriously than others?

In addition, the reproducibility of the study requires that the decisions taken are transparent, recorded and adhered to in a consistent manner, so that, in theory, another person using the same data and criteria would come up with the same result. This is especially true when carrying out clerical matching (by definition, these receive a human interpretation), but even on automatic matches lack of transparency can be problematic.

When the purpose of linking is for a specific piece of analysis, a statistical approach can reduce the effect of uncertain linkage; multiple imputation (using statistical models to fill in the gaps in the data) could be an acceptable alternative. In this approach, the 'uncertain' matches could most productively be used by treating them as the starting values for an imputation procedure. The probabilities from the match process would give an indication of how much weight to place on these starting values[11]. However, this approach is only relevant where linking and analysis are part of the same process, and does not assist in, for example, creating master keys for multiple linking.

Probabilistic linking is more tolerant, by design, of data errors then exact matching. Nevertheless, preparing the data for linking can require a substantial amount of data cleaning to remove 'filler' words and unhelpful terms; for example, "Mr John Smith esq" becomes a simple "JOHNSMITH". This does not deal with problems in the data itself caused by automatic systems having to recognise words. For example, to a human reader the town in "17, London Road, Birm., B16AS" is Birmingham, but this may not be recognised by a computer. As a result, string-matching algorithms are an ongoing research topic, and the choice of the algorithm can affect the outcome substantially. Sound-based matching would find that "Jon" is closer to "John" than "Beat" is to "Beath", whereas bigram analysis (splitting the text into pairs of letters, such as "be/ea/at" and "be/ea/at/th") would come to the opposite conclusion.

The matching of fields can require extensive computational resources. At its simplest, consider comparing two databases of M and N observations. Even for a single field, this requires MxN comparisons to be carried out. This can be made more efficient by sorting the fields and only searching the 'neighbourhood' of the target observation. However, this assumes that the data is

[10] For a review of some popular tools, see Tuoto T., Gould P., Seyb A., Cibella N., Scannapieco M. and Scanu M. (2014) Data Linking: A Common Project for Official Statistics. Paper presented to the 2014 Conference of European Statistics Stakeholders, Rome, November.
[11] Goldstein et al (2014)

observed without errors that significantly affect the order; this might be true for age, for example, but not for names.

One popular way to improve efficiency is by 'blocking' the text. This uses exact matching at a broad level where the link-maker is confident that data is accurately referenced across all data. For example, it may be a practical working assumption that the first part of a postcode is correct; even if this provides spurious accuracy to a small number of cases, this might be outweighed by the processing gains from having to match text of half the length. Alternatively, it might be decided to block on gender, on the basis that if there is a disagreement on such an important field the value of a link on other fields is minimal.

A1.2.3 Statistical linking and data fusion

Both exact and probabilistic matching aim to link the same individuals together, and they dominate practical projects. However, statistical techniques (sometimes called data fusion[12]) have been developed to allow analysis where the records of two different individuals have been linked as if they refer to the same person.

The premise is that if John Smith is a 42-year old white male surveyor, some of his characteristics of interest (such as education, earnings and political views) are likely to be similar to those of other 42-year-old white male surveyors. If this is the case, then linking medical data (for example) from the original John Smith to any one of these other similar candidates should give statistically similar outcomes.

The advantage of this method is that the quality of the link is less relevant as, by construction, any one individual in a group is much like another, but the method relies upon a number of strong statistical assumptions. Key is that the variables of interest in the two datasets are independent of each other, given the match variables. This is essential for the assumption that any one link candidate is as good as any other.

Statistical linking has been exploited by commercial organisations as a way of generating synthetic data for analysis which has some statistical basis. For example, a supermarket may have data from loyalty cards on a million customers, and may also have a small survey of a thousand customers, asking detailed questions. Typically the survey data is analysed by itself, to make inferences about the population. However, the supermarket could decide to fuse the survey data to its customer database via appropriate link fields, giving it a million pseudo-survey responses. The key is that this pseudo-survey reflects the actual distribution of customer characteristics, and allows unconnected variables to be analysed jointly. This sort of analysis requires additional assumptions about whether the survey represents the population of interest adequately, but within that it allows the organisation to draw links between variables with some statistical justification.

A major problem with statistical data linking is that the properties of any analysis are largely unknown if the assumptions about the data are not met. Papers propounding statistical linking typically take the perspective "*if* these conditions hold, *then* this is the result…" without formulating an alternative perspective. This is logical as the range of alternative outcomes is infinite, but not

[12] In computer science, 'data fusion' generally means what we have referred to as deterministic linkage

helpful for researchers when for genuine data the assumptions about the data appear to be somewhat heroic.

A1.2.4 Multilevel linking

Although data linking is normally thought of as between units of the same type (person-to-person, or organisation-to-organisation) there is potential in linking 'vertically' (individual to doctor to hospital, for example) or 'horizontally' between different dimensions (individual to small area data). For example, linking HIV infection with geographical data showed a substantial difference in infection on two sides of a river, which was not being identified from hospital admissions.

Multilevel linking produces fewer issues than other techniques. First, the match tends to be deterministic as the links to the higher level are in the individual data or not. Second, higher level data (for example, air pollution indicators) are more likely to be publicly available and so not subject to confidentiality constraints. This does not mean that data can be linked without restriction, as that public-but-linked information may help with identification of the detailed record. For example, if the air-pollution indicator has a unique value in one small area and is linked with confidential data on respiratory disease, that indicator would allow the small areas to be identified even if it is not included on the dataset.

A1.3 Characteristics of types of data

A1.3.1 Cross-sectional survey data

Surveys tend to be used to collect socio-economic data; the characteristics of the population, particularly where the data is less sensitive. As these are collected for statistical purposes, the data tends to be superficially clean – collected and produced to a common standard, with common definitions and ideally metadata. For government data collection, a substantial amount of time is typically spent on questionnaire design, so that there is clarity about the meaning of questions being asked.

The major problem with survey data collection is ensuring that it is representative of the population of interest. To try to keep survey costs down, techniques such as clustering (focusing on particular areas or groups) and stratification (using different sampling methods based upon some external characteristics) are used to focus effort on the most valuable observations. This even holds for census data where decisions need to be made about how much to chase up hard-to-reach respondents.

A problem considered less often is how accurate the data is. Sampling is an acknowledgement that not all data can be collected and, within certain parameters, one observation is as good as another. An error in determining someone's age is an unavoidable consequence of anything less than infinite resources, but statistically the expectation is that such errors should make no difference to analysis unless the errors are systematic and/or correlated with other variables in the dataset.

Survey data is also rarely checked for its accuracy once collected, as following up a survey respondent is likely to be expensive and may be impractical. Given the expected limited statistical impact, following up is rarely cost effective. There are some exceptions; for example, when the wage data collected for the UK minimum wage calculations shows a worker apparently being paid an unlawful wage, this is verified with the respondent. In other cases, survey forms can automatically

check inconsistent results (e.g. female suffering from testicular cancer, parent carer with no children under 18). However, in general the trend is reversing, with government statistical agencies (the main source of socio-economic survey data) increasingly using 'statistical editing'; that is, checking to see whether changing an unlikely result to a more reasonable one would affect aggregates, and only checking if this would be the case.

Whilst this data collection strategy is sensible for statistical organisations, for data linking this is potentially a problem. An age being recorded as 43 instead of 42 may make little difference to the analysis of that dataset, but it may prevent valid matches from taking place.

A1.3.2 Cohort studies and longitudinal studies

Cohort studies differ from cross-sectional surveys in that the subject is repeatedly interviewed; moreover, because re-interview is expected, the cohort planners will actively try to ensure that contact is maintained with the respondents after each wave of data has been collected. This provides additional checks for the quality of the data, as well as a mechanism for following up queries. If need be data can even be edited retrospectively.

Cohort studies have many advantageous statistical properties; their major drawback is the cost associated with managing a complex data collection operation where substituting one respondent with a statistically similar one is not an option. As a result, cohort studies tend to be much smaller than cross-sectional counterparts.

As far as linkage is concerned, cohort studies should be an easier proposition than cross-sectional studies as maintaining accurate identifying information is essential to keep the cohort going. Linkage can also pay dividends to the cohort. A major statistical problem is attrition; that is, people dropping out of the cohort. By definition, it is difficult for the cohort planners to know why someone leaves their cohort study, but linking with other studies may show that, for example, the individual has died, moved house or changed name.

A1.3.3 Register data

A number of countries maintain extensive registries of the population; some are general – for example, to manage ID card systems – but others may be specific to particular areas, such as health or education. The purpose of a register is to provide coverage of the population in question, and so these should be comprehensive and accurate data sources. This has great statistical potential as it reduces the problem of selection bias considerably, to whether the person is included in the register or not (in contrast, selective studies such as cohorts or surveys require both respondents and responses to be acquired effectively).

Ideally these registries use common personal identification numbers, which makes data linkage fast and accurate. Even if different IDs are used, registers are designed to be continually updated with new information. This means that the information needed to match is continually maintained and potentially available. The most extensive systems of general registers occur in the Nordic countries, although many countries hold registers for particular illness such as cancer.

A1.3.4 Other administrative data

The great advantage of administrative data (that is, data collected through normal operations) is that it can often be a census of the population of interest. Hence linking to administrative data can be done without reducing the number of cases for study; in fact, as noted above, it can provide both the study group and a control group, improving the robustness of findings significantly. This can compensate for the three main disadvantages of administrative data: semantics, quality, and variable range.

Administrative data is collected for operational needs, not statistical ones. Two general practitioners (GPs; primary care doctors in the UK) may record the same patient's illness differently depending on the perceptions of the patient's needs, history and prognosis. Guidelines may be unclear, may change over time, or may be subject to different interpretations. The GP's main interest is to ensure that the patient's medical notes make sense to him or her, not whether they are using an interpretation consistent with colleagues. Similarly, the range of variables in administrative data is determined by the operational needs of the business. This is not just organisations saving money: an organisation which routinely collected irrelevant data on its customers would be likely to face strong criticism. Hence GP data is unlikely to contain information on socio-economic variables such as income or detailed occupation, while tax data does not record ethnicity.

On quality, administrative data are likely to have been inputted by a large number of people over long periods of time; therefore the chance of data coding errors and inconsistencies, spelling mistakes and so forth is probably much higher. Moreover, errors in the data, even if discovered, would not necessarily be corrected. Administrative data is liable to be read, used and reviewed by humans who can interpret inconsistencies in the data correctly. In contrast, data linkage requires machine-readable consistency of data.

However, a number of researchers have challenged this perspective, particularly with respect to health; they note that, while all data is subject to error, administrative data input at the time it was needed is likely to be less error-prone than data collection methods relying on occasional updates or recall. This is because the data is needed for medical procedures, and any errors are likely to be identified quickly. On this view the argument about the accuracy is not about whether the data entered were correct, but whether the information known at the time was accurate (for example, a disease might not manifest itself immediately, or a patient in emergency care may be in no position to confirm personal details). The issue then is not whether the data collected are accurate (the argument would be that they are better than other data) but whether the most relevant data could be collected at all.

A2. The value of data linking

The ability to link different data sources together is crucial to epidemiology for a number of reasons. Broadly, these can be seen as: increasing the range of questions that can be asked; providing the historical perspective necessary for many studies; improving the statistical properties of any analysis; and making better use of resources.

A2.1 Increasing the range of feasible topic areas

A2.1.1 Identifying the correlation between health events from different sources

Health data may not be collected by the same organisation; even if they are, there may be separate registers for cancer, diabetes, genetic illness and so on. In addition, data collection for a clinical trial, for example, may be focused on addressing a specific research question. This ensures that the data collected is strictly necessary for the research, but may limit the opportunity to address slightly different research questions. Combining health data from multiple sources may allow interrelated effects to be investigated; for example:

- linking ambulance calls, emergency department data and hospital admission records to investigate pathways through health services for alcohol-related admissions[13];
- systematic reviews had highlighted the lack of sufficiently long clinical trials to evaluate the likelihood of cancer risk from insulin glargine; linking a diabetes register and a cancer register in Scotland was able to demonstrate robustly the lack of risk[14].

A2.1.2 Identifying contributory factors from non-health data

Health data focus on the specific event of the database (such as cancer progression) but may not have much data relating to the potential contributors (such as activity levels or family history). Combining health data with other data sources may allow the data to be broken down in different ways, and make it possible to answer questions which a single data set cannot resolve; for example:

- combining hospital records with immigration data from airports to analyse the incidence of deep vein thrombosis after long-haul flights[15];
- linking police arrest data and psychiatric records to evaluate how well mental health problems were identified at police stations[16].

In each case, none of the individual data sources were able to provide information on both the illness of interest and the contributory factors.

A2.1.3 Long term study

Health events can be experienced over an extended period, and tracking all relevant events over such a long period may not be feasible in a single database without excessive intrusion and/or cost. Using additional data which records such information as a matter of course (such as re-admission to hospital or prescription data) can improve the accuracy of data collection and reduce the burden on both observer and subject. For example:

- a Scandinavian register-based study of the impact of radiation therapy for cancer on the incidence of heart disease looked at up to forty years of health records for the female

[13] Matthews S., Ferris J. and Lloyd B. (2014) "Three datasets are better than one! Alcohol related diagnoses from ambulance to admission". Turning Point Alcohol and Drug Centre http://www.ihdlnconference2014.org/sites/default/files/GEORGIA%20B_APR29_1300_MATTHEWS.ppsx

[14] Colhoun H. and others (2009) "Use of insulin glargine and cancer incidence in Scotland: a study from the Scottish Diabetes Research Network Epidemiology Group". *Diabetologia*. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723678/

[15] Kelman CW and others (2003) "Deep vein thrombosis and air travel: record linkage study." *British Medical Journal.* http://www.bmj.com/content/327/7423/1072

[16] Baksheev,G., Thomas,S. and Ogloff,J. (2010) "Psychiatric disorders and unmet needs in Australian police cells" *Aust.N.Z.J.Psychiatry*

subject; this reflected the very long gestation period (hypothesised) for the radiation effects[17];

- one specialist in rare childhood disease suggested that just 75% were identified in childhood; the remainder took between five and thirty years to be identified, the rarity of the disease being the factor which stopped the illness being recognised[18].

A2.2 Providing the historical context or control

A2.2.1 Retrospective analysis

The effect of some conditions may not manifest themselves until many years after the initial incidence; alternatively, an illness may appear quickly but have contributory factors going back far into the patient's past. In both these circumstances, to study the illness it is necessary to have information going back to a period when there was no reason to collect information. Other than through prospective case control and cohort studies, this can only be addressed by linking health outcome data with information which was collected for other purposes, such as administrative data, vital events data, civil registration data or other sources.

While tracing back such information may be problematic, this has enormous statistical value. Because data were collected without reference to a particular illness, the inclusion of information from those who do not develop the condition can produce a ready-made control group; and as the data were collected in the past, the data are not subject to recall error. For example:

- in the Scandinavian study (cited above) on the effects of radiation therapy on heart disease, this was the first study to account for cardiac risk factors in the subjects at the time of radiation treatment rather than at the time of presenting with a cardiac condition

This can be very efficient for studies of rare health events. If an illness affects one in fifty thousand, then a prospective or case control study would need a very large number of observations to have a statistically useful number of cases. However, in a population of five million one would expect a hundred cases to be reported to the health service. These could form the treatment group (or subgroups), and analysts can concentrate on determining an appropriate control group.

A2.2.2 Prospective data collection

A parallel to the retrospective study is the prospective cohort study. This identifies a cohort of people and follows them over time, in more or less detail. As for retrospective analysis, the great statistical advantage is that groups are chosen before any medical conditions arise, and so 'baseline' information on all subjects can be collected before treatment and control groups are identified; again, data is collected at throughout the period and so recall error is not an issue. Prospective cohort studies can also focus on particular types of individuals to improve the efficiency of data collection (for example, focusing on a particular ethnic group which is susceptible to a particular disease). For example:

[17] Darby S. and others (2013) "Risk of Ischemic Heart Disease in Women after Radiotherapy for Breast Cancer". *New England Journal of Medicine*. http://www.nejm.org/doi/full/10.1056/NEJMoa1209825

[18] Van der Valk T. (2014) *A right to profit from research: patient perspective*. Presentation to CPDP 2015. https://www.youtube.com/watch?v=tRmODJ5lmDw

- the UK National Survey on Health and Development (also known as the 1946 Birth Cohort) has been providing microdata (more recently, linked microdata) to researchers for almost seventy years[19];
- a prospective study on the risk of hip fractures was able to distinguish the long-term impact of alternative exercise regimes and other factors by following a cohort of post-menopausal women for up to nine years[20].

The great disadvantage of prospective studies is the very large cost associated with recruiting and then following a large group of people (and drop-outs from the cohort are more likely to be from particularly groups than random attrition, potentially biasing results). However this large initial investment and ongoing expenditure is best leveraged by allowing such data to be linked for multiple uses.

A2.3 Improving the statistical basis

A2.3.1 Co-morbidity

Multiple health events can occur at the same time, or be associated with multiple concurrent socio-economic factors. These might not be recorded together as each data collection agency is focused on the outcomes most relevant to them. Bringing these records together allows co-morbidity to be investigated; for example:

- research at UCL used linked data to show that comorbidity may lead to significant under-reporting of (potentially preventable) deaths from respiratory tract infections[21];
- a study of hypertension showed that retrospective analysis of comorbidity before the diagnosis of hypertension improve mortality predictions significantly[22].

A2.3.2 Checking and improving data quality

All data contain errors to a greater or lesser degree. Combining multiple datasets allows the consistency of data to be checked, and potentially enables missing data to be filled in. For example:

- linking midwives data to vital events registers showed that previous estimates of births in one ethnic group had been misclassified to the dominant ethnic group[23];
- 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[24];
- 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[25].

---

[19] http://www.nshd.mrc.ac.uk/

[20] Armstrong M. and others (2011) "Body Mass Index and Physical Activity in Relation to the Incidence of Hip Fracture in Postmenopausal Women".*Journal of Bone and Mineral Research* http://onlinelibrary.wiley.com/doi/10.1002/jbmr.315/pdf

[21] Hardelid P., Dattani N., Cortina-Borja M. and Gilbert R. (2014) "Estimating excess winter deaths due to respiratory tract infections in children: a linked data approach"

[22] Chen G. (20140 "Influence of databases and look-back intervals to define comorbidity among newly diagnosed hypertension cases"

[23] Freemantle J. and Ritte R. (2014) "Using population data linkage to make the 'invisible' visible"

[24] Sveréus S., Dahlgren C., Brorsson H. and Rehnberg C. (2014) "Fooled by the means". http://www.ihdlnconference2014.org/sites/default/files/GEORGIA%20B_APR30_1030_SVEREUS.pdf

A2.3.3 Analysing rare events

By their nature, it is difficult to generate sufficient information on rare events from single data sources. Suppose that twenty hospitals each have a single incidence of a rare cancer; no hospitals can carry out a meaningful analysis using its own data, but pooling the data across hospitals may allow common features to be identified. For example:

- Marshall Smith Syndrome currently has approximately 23 sufferers worldwide; without data sharing, there is no effective analysis possible[26].

A2.3.4 Multilevel modelling

By combining personal data with information about groups, areas, systems and so on, it is possible to draw out contributory factors which reflect structures in society (including the structure of research groups). For example:

- William Farr and John Snow focused on drinking water delivery systems in their attempts to understand cholera in mid-19th century London, eventually demonstrating that a water-borne pathogen was the only feasible conclusion;
- in South Australia a Cancer Atlas was built to analyse, amongst other things, whether regional variations in access to care was affecting survival rates for geographically concentrated communities[27];
- a second Australian study used linked data to break down the multiple effects of locality, service provision and mode of transport in explaining differential traffic collision rates amongst ethnic groups[28];
- a multi-level study of caesarean section rates identified significant differences between hospitals and treatment groups, leading to a number of specific policy recommendations for improved practice[29].

A2.3.5 Generating useful tools

Single-source data are likely to be limited in their wider applicability. In contrast, linking data from multiple sources can allow population level tools to be developed. For example:

- linked data was used to generate improved modelling of diabetes risk factors in the Canadian population[30];

---

[25] Hoving J., Fritschi L., Benke G., McKenzie D. and Sim M. (2005) "Methodological issues in linking study participants to Australian cancer registries using different methods: lessons from a cohort study". Australian and New Zealand Journal of Public Health

[26] Van der Valk T. (2014) *A right to profit from research: patient perspective*. Presentation to CPDP 2015. https://www.youtube.com/watch?v=tRmODJ5lmDw

[27] Sharplin G., Bannister,S., Eckert M., Roder D. and Wilson B. (2014) "A South Australian Cancer Atlas shows important variations in cancer risk and outcomes, but can better use be made of Australian data to support the work of Cancer Councils?" *Cancer Forum*, July. http://www.cancerforum.org.au/Issues/2014/July/Articles/South_Australian_Cancer_Atlas.htm

[28] Jorm L. (2014) "Partitioning variation to explore outcomes"

[29] Lee YY, Roberts CL, Patterson JA, Simpson JM, Nicholl MC, Morris JM, et al. (2013) "Unexplained variation in hospital caesarean section rates". Med J Aust. 2013; 199(5): 348-53. https://www.mja.com.au/journal/2013/199/5/unexplained-variation-hospital-caesarean-section-rate

[30] Rosella L. (2014) "Risk Prediction with Linked Databases: The Diabetes Population Risk Tool (DPoRT)".

- Statistics Canada has developed a range of microsimulation models based on linked health and socioeconomic data to analyse policy impact and the robustness of health management systems to unexpected events[31],[32].

A2.4 Improved use of scarce resources

A2.4.1 Making data analysis more timely

Linking data from existing sources for analysis may well be the quickest way to get the answer to a statistical problem. Although getting approval for access to the data may take time (as might learning about the data), there is no additional time to collect the data, and so analysis can be achieved relatively swiftly. For example:

- when concerns were raised about the lack of evidence on carcinogenicity of insulin glargine, following a change in official recommendations, a Scottish study was able to provide a comprehensive response within six months from linking cancer and diabetes registers[33];
- a UK study suggested that better use of existing data in live analysis could create savings of around £1bn per annum on the NHS budget (top end estimates) by reducing the number and severity of health incidents[34].

A2.4.2 Cost

Dedicated data collection is expensive, particularly from medical sources. If that data can be re-used then the public benefit can be substantial. For example:

- a Swedish study in the 1990s analysing the potential carcinogenic effect of vitamin injections in children took just three months to compete and required no new data collection; all the information was already held in the registers and was accessible to the research team[35];
- an Australian study of vitamin-D deficiency using existing cohort data provided the *prima facie* evidence for a more targeted case-control study[36].

A2.4.3 International comparisons

Sharing or linking data between countries is often difficult, because transferring identifiable data out of countries is typically more difficult than sharing within the country of origin. Nevertheless, international data sharing and linking, if feasible provides a number of benefits:

- for rare diseases, this may be the only way to get sufficient observations to allow analysis;
- linked analysis (not data) across countries can help to identify specific cultural or regional factors, as in the Alpha Network[37] or INDEPTH[38] projects (see case studies).

---

[31] Wolfson M. (2014) "Answering Questions that Matter: from Data Linkage to Microsimulation Modeling"

[32] For a review of microsimulation in health generally, see Zucchelli E., Jones A. and Rice N. (2012) "The evaluation of health policies through dynamic microsimulation methods". *International Journal Of Microsimulation*

[33] Colhoun H. and others (2009), ibid.

[34] Volterra Ltd (2014) *Sustaining universal health care in the UK: making better use of information*. http://volterra.co.uk/wp-content/uploads/2014/09/Final-EMC-Volterra-Healthcare-report-web-version.pdf

[35] BMJ 2013 K-vitamin inject ion Magnus Stenbeck

[36] Wong YY, McCaul KA, Yeap BB, Hankey GJ, Flicker L. (2013) "Low vitamin D status is an independent predictor of increased frailty and all-cause mortality in older men: the Health In Men Study". *J Clin Endocrinol Metab*. v98:9 pp3821-8. http://press.endocrine.org/doi/full/10.1210/jc.2013-1702

[37] http://alpha.lshtm.ac.uk/

A2.4.4 Interdisciplinary research benefits

Finally, one advantage of sharing data from different disciplines is that it may encourage interdisciplinary research. As epidemiology explicitly recognises that the health of the public can be determined as much by socio-economic factors as by viruses or bacteria, an inter-disciplinary research environment might be more successful at identifying causes and effects, compared to social scientists, operational researchers, psychologists, clinicians and others operating within their own research disciplines. However, it seems an open question as to whether interdisciplinary working stimulates the development of new interdisciplinary data sources, or vice-versa.

A3. Problems of linking data

In theory, a researcher wanting to link data sources can call on many statistical and practical resources. The methodology of data linking is well established, as are the statistical pitfalls of linked datasets and the conditions necessary for analysis to be valid. For implementation, commercial and publicly available tools support deterministic and probabilistic matching, and 'trusted third parties' offer secure linking. Finally, the last ten years has seen a significant growth in the legal and technical framework around the management of confidential research data, particularly in the provision of general-purpose research data centres (RDCs; also called data enclaves in the US). While the medical profession has made use of physical RDCs for a long time, the new preponderance of 'remote' RDCs accessible from a range of geographical locations has revolutionised the use of confidential social science data for research. This increased availability of identifiable data in a secure research environment means that researchers are no longer restricted to anonymised data.

In practice, data linking is much less straightforward. Barriers to effective data linkage can be statistical, technical and/or institutional:

- *Statistical* barriers include: lack of data; missing or poor quality match fields; biased data collection; inappropriate assumptions (such as the independence between variables of interest and match variables); lack of control groups in administrative data; and inconsistencies in the timing of data collection.
- *Technical* barriers include: lack of access to appropriate secure facilities; difficulties in extracting data from administrative systems; restrictions on data flows; limitations on the persistence and ownership of a linked dataset; the effectiveness of matching algorithms; and practical issues arising from different IT systems and data processing standards.
- *Institutional* barriers include: legal limits; custom and procedure, particularly when misinterpreted as legal strictures; organisational culture, inertia and beliefs; trust in (government) institutions; poor communication/relationships between data holding agencies; lack of incentives to improve data access; and the lack of effective champions.

This is not an exclusive list, nor do all issues relate to all cases of data sharing in all countries. For example, in a low-income country dominated by ad hoc interventions to address specific medical emergencies, the lack of data is likely to be the biggest hurdle. In contrast, a high-income country with an integrated health service may find that institutional barriers to data sharing, both within and outside the organisation, are of most concern.

[38] http://www.indepth-network.org/

The difficulties also vary with project scale. A one-off project linking intervention data with a survey presents very different problems to a project trying to broker a permanent data-sharing arrangement between an integrated health service and a research institution.

This section concentrates on the difficulties found when linking data in practice, providing a brief summary of some of the issues. This is not intended to be an exhaustive review: the choice of material is selective, to provide the background for the discussion of findings from the interviews and case studies. The review is organised around the three topics noted above: statistical issues, operational/technical issues, and institutional ones.

A3.1 Statistical issues

Whilst all research data has some limitations, linking data generates a specific additional set of problems.

A3.1.1 Quality of the match fields

When analysing a single dataset, some measurement error can be tolerated; for example, age being recorded at 44 instead of 42 may have little effect on multivariate analysis. In contrast, this small variation may be sufficient to prevent links being made. Whilst modern software for linking can be made fault-tolerant (for example, recognising "Jon Smith" as "John Smith") each discrepancy casts doubt on the linking and so lowers the probability of a correct match.

A3.1.2 Consistency

Consistency of definition amongst match fields is important. For example, in public health, typical match fields (where there is no match variable such as health service number, for example) would be date of birth, gender, socioeconomic status and ethnicity. Both of the latter can be problematic: they may be difficult to identify, and their definitions may change over time

Where there is a potential hierarchy in the categories this can be managed; for example if one dataset stores ages as actual values but another as only five-year ranges, it is possible to convert without error from the more detailed variable to the latter. However, as the latter has fewer categories, it is likely to produce more multiple matches in a probabilistic linkage.

Where match field definitions are not hierarchical, collapsing categories is not feasible; this limits the scope for linking even if the smaller category is acceptable for analytical purposes. When Statistics NZ expanded its definition of ethnicity from (broadly) "European" or "Maori" to include "New Zealander", both European-descended and Maori-descended began describing themselves as "New Zealanders". For those individuals, ethnicity can no longer be linked to the earlier definition[39].

A3.1.3 Characteristics of the matched sample

If the likelihood of a good match is related to the characteristics of the individual, this will affect the quality of the match data. For example, if one of the dataset was a survey on drug addicts, it would be reasonable to expect that the most accurate information would be supplied by those with the most stable lifestyles. Hence the matched dataset is more likely to be missing out on chaotic drug

[39] Callister P. (2004) *Seeking an ethnic identity: Is "New Zealander" a valid ethnic group?*. Callister Group

users. This does not necessarily imply bias in studies, as bias is a function of the analysis. A dataset which is perfectly acceptable in one use may lead to biased inferences in another use.

The general perception is that linking can only reduce the representativeness of the study group for statistical analysis.: the linking process creates a dataset which is at most the size of the smaller of the two sources (with the exception of fusion models creating synthetic data), and which represents the combined sampling characteristics of both data sources. A linked dataset cannot be more representative of the study population than the source data, and if the match rate is less than 100% it will be less representative.

However, in public health data this situation is often reversed, because typically one of the datasets is a census of the population under review. For example, when linking a register of stroke victims with a survey of elderly patients, the non-appearance of some survey respondents in the register is an indication of absence of (diagnosed) stroke; therefore, a control and treatment group is immediately distinguished.

A3.1.4 Quality of the overall match

A key unknowable in data linking is the overall accuracy of the match fields (and, to a lesser extent, completeness). Research studies tend to concentrate on demonstrating the advantages of one linking technique over another on synthetic datasets, because then the modelled properties can be compared with the true properties; hence there is a lack of evidence from real-world cases to know how effective the claims from research really are. The difficulty is that any such study would necessarily be specific to a particular set of datasets, and would require knowing the true exact matches. Such a data set is unlikely to be representative of real, messy data; but even if it were, building up a picture of how important the match technique is would require the same assessment to be carried out on a wide variety of datasets. It is not clear who would have the data, expertise and funding to carry out such a study.

A3.2 Technical and operational aspects of data linking

A3.2.1 Acquiring permission to link

The legal aspect of acquiring data from different sources can be complicated by differences in:

- the authority to share data (for example the health authority and the Census office);
- the status of the data(for example, sexuality, ethnicity, health are formally classified as ‘more sensitive’ in European regulations);
- consent to link and use the data;
- organisational attitudes to risk, utility and lawful authority;
- approval processes, such ethics committees.

As well as differences between organisations, questions arise over the status of the data:

- Who will control the data?
- Does linked data count as ‘new’ data with a new data collector?
- How long will the linked data be used for?

Where the data are being linked for a specific research project, these are relatively straightforward. However, these can be potentially fatal stumbling blocks for projects to create a new linked dataset

archived for further research use. Consider a project to create linked medical records to Census data to provide a resource for further analyses by third party researchers. Amongst the questions the research data manager is unable to answer are:

- Who will use the data?
- What will the data be used for?
- How long will the data be needed?
- How can we know the users can be trusted with the data?

The research data manager may seek to persuade the data depositors that appropriate procedures are in place to ensure that data use is lawful and ethical, but this means that the body granting agreement to the linking is effectively agreeing to delegate some of its authority. This may be harder to sell than allowing the original data depositors to retain control over use of the data. However, if the aim of setting up the project is to improve the efficiency of data access, then going back to the original depositors of the data for each research use may not be practical.

A3.2.2 Agreeing the hosting protocol

Once approval has been granted, data needs to be transmitted to the research data managers. Best practice in linking data is to separate identifiers and variables of interest, so that only those who need to see are given any information:

Step 1: data depositors extract identifiers from the dataset

Step 2: data depositors pass identifiers to data linkers

Step 3: data linkers carries out link and generate non-identifying reference

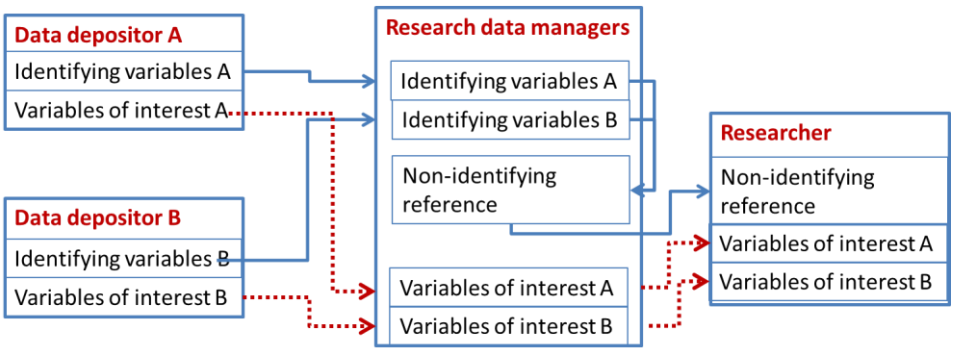
Step 4: reference is returned to data depositors along with identifying variables

Step 5: data depositors replace identifying variables with non-identifying reference and pass to research data manager

Step 6: variables of interest with non-identifying reference are passed to researcher

One model is that all data is transferred to the research data managers team (that is, they take on the role of data owners A and B, above), who then carry out the linkage; see Figure 1.

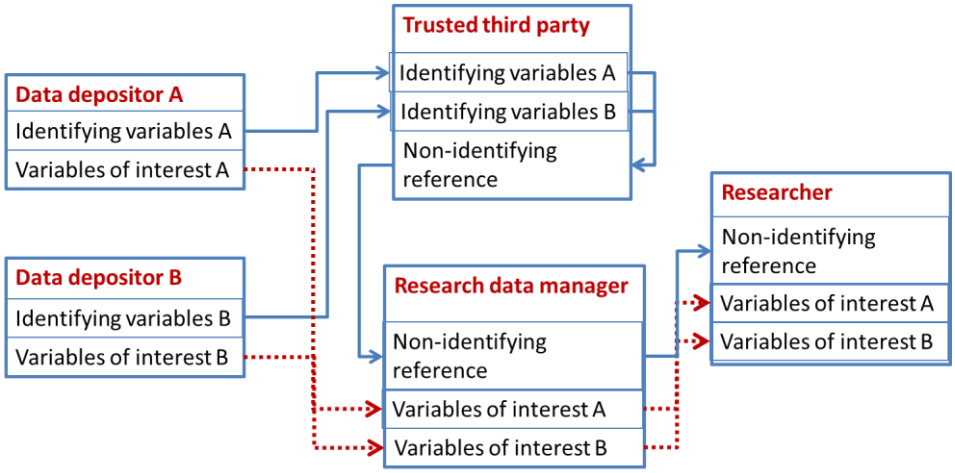
Figure 1 Project team as the linker



Even though the research data manager has all the data, separating variables of interest and identifiers is still seen as good practice, as it lowers the risk of accidental breach of confidentiality.

Giving both identifiers and variables of interest to the research data manager increases the amount of identifiable, confidential information out of the direct control of the data depositors. Hence, some linking occurs through third parties. The role of the third party is to ensure that no group ever has both identifying information and confidential data from any other party. In Figure 2 a trusted third party (TTP) is used; that is, the TTP is trusted enough to see the original identifying variables (note: the non-identifying reference should be returned by the TTP to the data owners so that it can be attached to the data, but this is omitted for clarity).

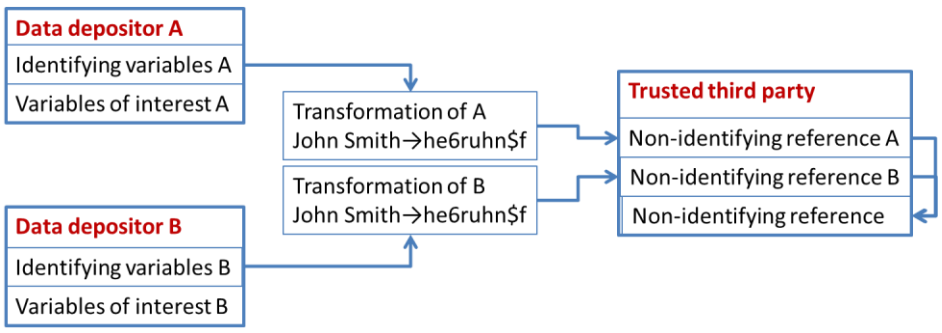
Figure 2 Trusted third party as the linker



The TTP model is easily understood and relatively widespread; there are dedicated organisations in both the public and private sector that offer TTP services, and this is the typical model of health data linkage in developed economies. Note that some research data managers may use trusted third parties even when they are one of the original data owners and they have just set up internal mechanisms to separate the processes. In this case the research data manager clearly has potential access to the complete set of identified data. However, the purpose of using a TTP is to demonstrate that the research data managers are making an additional effort to guard against casual identification; a self-denying ordinance to burnish their credibility.

In some cases, even this arrangement is deemed too sensitive, as identifying information is leaving the direct control of the data owner. Hence, the data depositor may use an untrusted third party (UTP). For this, the identifiers are transformed by a known (but irreversible) process into non-informative identifiers, which are then passed to the third party, and the process continues as before; see Figure 3 (detail removed for clarity).

Figure 3 Untrusted third party as linker



This often referred to as privacy-preserving record linkage (PPRL) as it avoids any directly identifying information leaving the data owner's direct control. For a privacy perspective this is very appealing, but it has significant practical drawbacks. The most obvious is that, in its simplest state, probabilistic linkage is not feasible. In probabilistic linkage, the linker needs to be able to determine that "JohnSmith" might be "JBSmith"; however, the point of PPRL is that, while "John Smith" translates into "he6ruhn$f", "JB Smith" translates into "kh67£G*aq" or something else suitably non-informative about the source data; otherwise, anonymity would not be preserved.

One solution is to apply some of the probabilistic linking techniques before the anonymisation is applied. For example, if the bigram technique[40] is being used then the transformed bigrams for the "smith" part of the name would still be comparable. However, this does not work where the whole field value is needed; for example ages 42 and 43 must generate unrelated transformed values, or the non-identifiability of the transformed data is no longer unidentifiable.

There are further potential problems. First, splitting the identifier into sub values before transformation increases the risk of the transformation process being undone via statistical analysis of repeated combinations, in the same way that simple replacement ciphers are broken. Second, the need to ensure the same transformation is applied requires sharing information about the transformation process. Third, the source identifiers must have been cleaned in the same way by both data owners. Finally, it is not possibly to carry out clerical matching on 'uncertain' matches; therefore the subjective choice of success parameters becomes all-important.

Nevertheless, this 'privacy-preserving probabilistic record linkage' (PPPRL) has attracted a lot of interest. One way forward may be to combine both elements – passing over less identifying variables such as age and gender, but anonymising address information.

Distributing data

An alternative to linking the data is distributed processing: allowing researchers to use the data but without directly linking it. Supposing a researcher has access to the non-identifying references of A and B. For some analyses, it might be possible for the researcher to send statistical commands to the data depositor A of the form "give me the value of x for individual he6ruhn$f". This value is returned with noise generated so that the actual value is not known. When the values have been collected from all individuals from all data sources, the generated statistics are then returned to the data depositors to remove noise from the aggregate statistic, leaving the true values exposed but in aggregate non-disclosive form[41].

The advantage to data depositors is that they can be sure that they always retain control of the data. The difficulty, apart from ensuring that sufficient observations exist, is defining a useful set of statistics to which this can be applied. It works well for univariate statistics, and it can be applied to simple linear regressions (which effectively involves repeated addition or multiple passes of the data) but is less valuable for more complicated observations where the interaction between individuals is important. Hence at present this is a relatively specialist application.

[40] Bigram matching separates words into paired letter combination (for example "Lesley" produces the bigrams "le/es/sl/le/ey", and "Leslie" leads to "le/es/sl/li/ie". See section A1 for more detail.
[41] For an example for health data, see http://www.datashield.org/.

A3.2.3 Acquiring the data

As noted above, data acquired from statistical sources tends to go through extensive and well-documented cleaning processes to produce a clean dataset, whereas administrative data are more likely to have semantic and quality problems. In terms of acquiring the data, survey data are also easier to deal with, as it is designed to be analysed statistically (although, for example, there are many different competing metadata standards).

In contrast, administrative processes create a number of 'syntactic' problems in the way that data are processed. Administrative systems designed for case-by-case operational use may not be able to produce whole datasets. Data held in analytical statistical systems (SAS, SPSS, R, Stata) can be readily transferred, often with the metadata as well; in contrast, extracting a meaningful file from an enterprise resource planning system with useful metadata may be a lengthy and unrepeatable process. The Wellcome Trust report on data discoverability[42] noted that documentation of datasets from different sources can be a significant barrier to effective use of linked data.

Finally, survey data have a clear start and end date, as does field data collection. Administrative data systems are more likely to be updated on a continuous and open-ended basis. Hence data collected from administrative systems are expected to change over time; repeated request for data may generate different outcomes.

A3.2.4 Providing access to researchers

Managing research access to linked datasets is possibly the least troublesome aspect of data linkage. Although linked datasets may be more sensitive than either of the source datasets individually, the landscape of data access has changed considerably in the last decade or so. Data managers can essentially pull a data access solution off the shelf[43]:

- *Anonymisation* to reduce the information content (and so risk) of data has a research history going back fifty years;
- *Licensing* of researchers, sometimes combined with a degree of anonymisation, is still the most common way for researchers to get access to data;
- *Secure 'research data centres'* (RDCs, also sometimes referred to as 'safe havens'), laboratory facilities with very detailed data but some physical restrictions or oversight;
- *Remote access*, sets up 'virtual' RDCs allow users to manipulate data unhindered by geography; implementation varies greatly from restricted-site access only to direct access from the internet;
- *Remote job submission* allows users to send statistical programmes to be run and return results; this is relatively uncommon but a few operations have adopted this route.

All of these have sufficient track records to be considered 'mature' approaches to user needs. Of course, specific implementations vary and each has its own characteristics in terms of whether trust is embodied in users, IT, legal consequences and so on. Major data providers, such as national statistical organisations, employ a number of these options, and often in combination.

[42] http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTP054675.htm

[43] Ritchie F. (2014) "Access to sensitive data: satisfying objectives, not constraints", *J. Official Statistics*

In recent years, there has also been interest in s*ynthetic data:* data which are expected to have the same characteristics as the real data but which are imputed from statistical models; the resulting dataset is then intended to be safe for distribution. There is no disclosure risk from invented data, but there is also no value in purely invented data. Synthetic data models hence use the source data characteristics, and may also mix real and synthetic data to make the synthesised dataset more realistic. The risk is that using more source material makes the synthetic data closer to the original, and so creating a potential disclosure risk. In public health the value of synthetic data seems low given the importance of accurately assessing recording health events; however, synthetic data have been used in data fusion models to generate simulation models for policy analysis.

The last decade has also seen an increasing formalisation in ways to describe, design and present data access solutions. For example, the 'five safes' model or some variant is widely used in HICs (particularly in the UK) to provide a common frame of reference for access discussions[44]; the Organisation for Economic Co-operation and Development developed a model of 'Circles of Trust' to improve international data sharing decisions[45]; and both of these are compatible with 'zoning models' such as that used for the TRANSfoRm project[46]. Although these access models use different terminology, the common feature of all is the recognition that data access is achieved through a balance of approaches; many different solutions are compatible with safe access, and the key decision are about costs and benefits, not about whether something is possible or not.

A3.2.5 Using linked data in research

Researchers, on the whole, have relatively little interest in where the data they use comes from. Nevertheless, most data depositors do provide some form of support to researchers, even if only a willingness to answer questions.

For linked data, the question of who should be providing this support arises. Each of the data depositors can be assumed to know their own data well, but are they as well equipped to advise on a linked dataset (or the quality of that link)? In addition, metadata is likely to reflect the interests of the data collecting organisations, not necessarily the research data managers.

A popular solution is to make the research data managers the new gatekeepers of knowledge. This appears to be an extra cost – more support staff are needed – but overall having an intermediary who can talk to both the data depositors and the researchers can be a cost-effective solution. If data depositors are not familiar with research methods then they might be overwhelmed by unexpected questions from researchers, and so an 'expert questioner' can develop a productive relationship; meanwhile, some of the time typically spent supporting new researchers can be gained back by having dedicated data experts in the research data team. Note however, that while this might be

[44] The 'Five Safes' framework proposes considering data access as a series of separate but interconnected decisions on project purpose, people, technical setting, data detail, and type of output; see Desai T., Ritchie F. and Welpton R. (2014) *Five Safes: designing data access for research*, mimeo, UK Data Archive

[45] OECD (2014) *OECD Expert Group For International Collaboration On Microdata Access: Final Report*. Organization for Economic Co-operation and Development, Paris, July. http://www.oecd.org/std/microdata-access-final-report-OECD-2014.pdf

[46] Wolfgang Kuchinkea, W., Ohmanna C., Verheijb R., van Veenc E., Arvanitisd T., Taweele A. and Delaneye B. (2014) "A standardised graphic method for describing data privacy frameworks in primary care research using a flexible zone model". *International Journal of Medical Informatics* v83:12 pp941–957 http://www.sciencedirect.com/science/article/pii/S1386505614001634; see also the TRANSfoRm website at http://www.transformproject.eu/ .

cost-effective overall, the *observed* costs are likely to appear as a cost of running the data linkage project (as opposed to the unobserved avoided costs to the data depositors).

A3.3 Institutional aspects of data linking

A3.3.1 Legal issues

Legal gateways generally require the identification of:

- who will use the data?
- under what authority?
- for what purpose?
- for how long?
- what will be done with the data afterwards?

Not all data legislation specifies how data can be used. For example, the 1948 Statistics of Trade Act which governs much data collection by the UK Office for National Statistics (ONS) imposes strict restrictions on who can re-use the data, but the 1921 Census Act makes almost any statistical analysis potentially lawful should ONS agree to it. The Walport-Thomas Review[47] noted "the law itself does not provide a barrier to the sharing of personal data. However, the complexity of the law, amplified by a plethora of guidance, leaves those who may wish to share data in a fog of confusion." (Foreword, para 4).

However, modern laws generally require one of two approaches: consent, or a specific gateway relating to research access.

Consent

A person consenting for his or her confidential data to be linked and analysed is often referred to as the 'gold standard' gateway[48]. It provides both an ethical and a legal framework for managing and using data. However, it may be both impractical and undesirable, and the process of gaining consent itself may cause ethical concerns.

First, there may be the difficulty of contacting the individual who has moved away, for example. If the individual has died, consent is clearly impossible but data protection laws might still pertain to the use of that data (for example, it might affect other members of the family).

Second, the scale of gaining consent may also be impossible: sending out many thousands of consent forms may make the costs of the project unworkable.

Third, gaining consent may be undesirable as it breaches confidentiality. All of the UK Census-based longitudinal studies use sampling mechanisms based upon birth dates. Contacting an individual for consent reveals that that person is a candidate for inclusion, and so his or her birthdate increases the likelihood of identification of others. As another example, consent to release DNA information might lead to the (unconsented) release of DNA information of close relatives.

[47] Thomas R. And Walport M. (2008) *Data Sharing Review Report*. Health and Social Care Information Centre. http://systems.hscic.gov.uk/infogov/links/datasharingreview.pdf/view

[48] See, for example, Brosnam T, Perry M. (2009) "Informed" consent in adult patients: can we achieve a gold standard? *Br J Oral Maxillofac Surg*. Apr;47(3):186-90.

Fourth, consent may lead to biased samples. Studies on the propensity to give consent to link tend to show that there is a difference between those agreeing to the linking and those disagreeing. This does not necessarily bias analysis but it gives cause for concern.

Fifth, even if consent to use data is given, this may still result in biased samples. For example, it is straightforward to show in principle that in cases of terminal illness the delay caused by waiting to gain consent can bias outcomes[49].

Finally, the decision to give consent is very likely to be influenced by the surroundings, the interviewer, the way the question is phrased and so on[50]. 'Giving consent' is not an objective decision.

Implicit versus explicit consent

The accepted practice is that consent should be informed and explicit; that is, individuals should know what is being done with their data. This may not be an easy message to convey[51]. How explicit should a consent form be? If it is too detailed, it might unnecessarily restrict the use of the data to a very specific piece of research[52]. Moreover, a detailed explanation to a non-technical audience may lead to box-ticking: accepting terms and conditions without having read or understood them, in much the same way that we tend to accept software licences, for example. Courts have shown themselves willing to argue that clauses which the consumer could not be reasonably expected to read and understand may not be enforceable as the 'consent' is not informed.

Hence consent forms are more likely to state the use of data in more general terms: "this will be used for research purposes only in accordance with NHS guidelines; no data that identifies you will be distributed". Such a statement places the responsibility for ensuring legal and ethical compliance with the experts defining and monitoring the guidelines. However, so -called "broad consent" may not be within the spirit or letter of the relevant laws.

Opt-in vs opt-out

In recent years proponents of behavioural psychology have argued that significant changes in behaviour can be brought about by small changes in the way things are perceived or actions need to be taken. For example, in Denmark and the UK, roughly the same proportion of individuals choose to opt out of the default organ donor arrangements, despite the default option being "donate" in Denmark and "do not donate" in the UK; other countries show similar findings. This has led to calls for improving consent rates by requiring users to opt out rather than opting in.

Ethically, this is problematic. If few individuals reject the default option, this suggests they are not making a conscious choice; so how can they be genuinely said to 'consent'? On the other hand, if most individuals accept the default, who is to say they have not considered the issue, and come to the conclusion that the default is at least as good as any other option? Given the known importance of social norms in individual decision-making, it could be argued that the idea of 'consent' as rational information-driven decision-making is overly idealistic.

---

[49] Rookus M. (2014) "Narrow informed consent and observational medical research". Presented at CPDP 2015. https://www.youtube.com/watch?v=UDZwFjrqNj0

[50] See for example Korbmacher J. and Schroeder M. (2013) "Consent when Linking Survey Data with Administrative Records: The Role of the Interviewer" *Survey Research Methods* Vol.7, No.2, pp. 115-131

[51] Brosnam T. and Perry M., ibid

[52] Rookus M. (2014), ibid.

Competing jurisdictions

Even if the legal framework is clearly defined, projects may suffer from needing the approval of multiple jurisdictions. For example, The Menzies Foundation identifies Australia's federal health care system as the largest single barrier to effective research use of Australian health service data[53]. This is not because the states have different laws (although interpretations might differ), but because each state believes it cannot delegate its legal obligation to review uses of its data. Similar complaints have been made by German researchers, who also face an exceptionally decentralised system; or by any project where more than one ethical committee feels it has jurisdiction.

It may be worth considering what is being argued over. Each body believes it has a legal requirement to exercise jurisdiction. That may be true, but it does not follow that each body must carry out its own enquiry into the application. Refusing to accept the judgment of another body implies that you believe the other body is not competent to decide on what makes a scientifically valid project with adequate ethical guarantees and confidentiality protection. On the other hand, accepting the recommendation of another body that a proposal be approved does not mean that you are ignoring your legal duty, just that you are convinced by the evidence supplied by that other body that due diligence has been carried out. In other words, a one-speaks–for-all solution is lawful, with some tweaking of the approvals process.

Given that some of the most interesting developments in public health are the relationship between medical and socioeconomic factors, competing jurisdictions for approval are likely to be a barrier.

Use of research gateways

Because consent is not problem-free, many countries have a legally mandated (statutory or common law) gateway allowing access to data for research purposes. In Europe, the current data protection regulations allow research use of data of personal data for observational studies; in the UK, access to medical data is explicitly authorised in primary health legislation. Such legislation typically also specifies that there be appropriate checks and balances to ensure that data collection is consistent with the spirit as well as the letter of the law.

This provides a different set of challenges. Because the agreement of the individual is not required, research gateways may be perceived as something underhand, smacking of Big Brother. Research gateways may also require more explanation than the relatively simple concept of consent, particularly if the data being linked is covered by different legislation.

Law versus custom

Law is rarely a black-and-white issue; the legal system is a recognition of the fact that general laws need to be interpreted and understood in specific contexts. One consequence of this dichotomy, of the (perceived) status of law as unambiguous rule and the (actual) practice of law within context, is a tendency for a well-established custom or procedure to be mis-interpreted as a legal requirement.

---

[53] Menzies Foundation. (2013). Public Support of Data Linkage for Better Health. Available: http://www.menziesfoundation.org.au/pdf/Data%20Linkage_16aug13/Menzies%20Foundation_Public%20support%20for%20data-based%20research.pdf

This is most likely to occur where, in the absence of explicit legal statements, institutions are tasked with deciding the interpretation of the legal framework[54].

This may affect the ability of research gateways to operate, particularly if processes have evolved over a long period of time. For example, in one country, access to Census data is still largely discussed in terms of legislation several decades old, despite a more modern law being in place which reflects modern research needs. Research gateways can suffer from 'regulatory capture' by institutions keen to ensure that their interpretation of law prevails.

Defining confidentiality

A final legal issue concerns the interpretation of 'confidentiality'. Whilst legislation may use such terms as 'confidential' and 'anonymised', there is no legal definition. Instead it is left open for a competent authority to determine, and/or reference to be taken to 'reasonableness'. This is the case for the current European data protection regulation, which explicitly has a 'reasonableness' test. In rare cases a law specifies a minimum frequency, for example, but these are always predicated on the assumption that the data are 'confidential' – which takes as back to the beginning.

Generally the uncertainty allows research use, but occasionally this can have the opposite effect. The Australian Statistics Determination 1983[55] blocks access to data unless the manner of access "is not likely to enable the identification of [the subject]" (section7(1)b). As any release is a precondition for "enabling" identification, then so too any release is not just likely but certain to enable identification, even if the chance of a successful identification is tiny. A strict interpretation of the law bans all data release, including aggregate statistics, which is clearly nonsensical.

Hence, a key part of the legal framework is left open to human interpretation. This is sensible as confidentiality is always specific to the context, and so it is unlikely that any meaningful context-free definition could be written in case law or statute. Nevertheless, it is worth recalling that, like the interpretation of law itself, confidentiality is a human construct; even if the decision is primarily technical, subjective perspectives on risk, evidence and the philosophy of data access can generate quite different outcomes. Thus two organisations considering the confidentiality of a linked data source can come to different conclusions, each consistent with the data depositor's perspective.

A3.3.2 Ethical concerns[56]

Law and ethics are interrelated. Law can be seen as general rules of ethics enshrined in text; research ethics committees (RECs) are the interpretation of the spirit of the law in specific contexts. Even more than confidentiality, the ethical perspective is a human artefact, and liable to vary from organisation to organisation. The additional difficulty is that, while confidentiality assessment is primarily technical, ethical assessment requires balancing competing subjective claims: the rights of the individual against the rights of society.

---

[54] Ritchie F. (2014) , ibid.

[55] http://www.comlaw.gov.au/Details/F2004C00203

[56] For an extended debate on this topic, see the two sessions "A learning health care system: secondary use of data in research" https://www.youtube.com/watch?v=UDZwFjrqNj0 and https://www.youtube.com/watch?v=tRmODJ5lmDw ; or for a specific review see Meslin E. (2013) *Navigating the Policy "Valley of Death" in the Data Linkage Debates:Getting the Ethics Right.* Menzies Foundation Presentation. For a comprehensive investigation into ethical policices ofr a national health care system, see Nuffied Council on Bioethics (2014), *The collection, linking and use of data in biomedical research and health care: ethical issues* http://nuffieldbioethics.org/wp-content/uploads/Biological_and_health_data_web.pdf

Ignoring the statistical issues associated with the need to gain consent, the standard starting points for debate are that:

- the individual has a right to privacy and therefore control over his or her data (i.e. informed consent must be present);
- the government has a duty to act in the interests of society as a whole and may override the wishes of an individual (i.e. informed consent cannot be insisted upon).

The first point is found in numerous documents which start from the premise of the 'Nuremberg code', that voluntary participation is essential. However, the argument that consent is neither necessary nor sufficient to prevent harm is easily demonstrated: the absence of harm in (non-consensual) observational studies is the overwhelming case in research studies, whereas the Tuskegee syphilis study[57] or the clinical trials for TGN1412[58] showed that consent does not protect from harm.

The cases for overriding the need for consent can be described as paternalism, self-interest, and solidarity or reciprocity.

The paternalist case is that the state has more and better information and can therefore make a more informed judgment than the data subject. Whilst behavioural psychology has demonstrated that humans are very poor at processing complex decisions, the evidence that state planners are better is not clear; more importantly, this argument can be used to justify a range of undesired behaviours by the state, and so the paternalist argument is now used rarely, if at all.

The self-interest argument is based on the uncertainty of research. If your data are used, there may be a small loss of privacy to you, and the gain to society's knowledge may not benefit you. On the other hand, there may be a gain to you at no loss, because someone else's privacy has been reduced (by their data being used). As research is uncertain, it is impossible to know whether research will benefit you or not; however, it is clear that if no research is carried out, all will be worse off. Therefore, given that there is ample evidence that privacy can be managed by the research community, it is in everyone's own interest that their data be used for research.

This argument is often used to persuade participants to give consent to their data being used for research. However, both for gaining consent and for over-riding consent, the obvious problem of this cost-benefit argument is that sometimes the cost definitely exceeds the potential benefit: for example, taking tissue samples from elderly men to study childhood diseases or ovarian cancer. Hence, an extension of this is the solidarity/reciprocity argument, which seems to be the most popular argument at present.

Like the self-interest argument, the solidarity/reciprocity argument uses the fact that research is uncertain. Unlike self-interest, the argument here is that there is no connection between costs and benefit. Instead an individual is seen as part of society; sometimes he or she bears costs for the good of society, and sometimes he or she receives the benefit. This is part of the social contract: I help with 'research' without knowing who I help, because others help me without knowing it. Hence, even if I can see no benefit (females whose participation is deemed useful in a prostate cancer study,

[57] http://www.cdc.gov/tuskegee/timeline.htm
[58] http://www.i-sis.org.uk/LDTC.php

say) I should be willing to participate. Unlike the self-interest case, this does produce a moral argument for participation in research.

It also produces a case for over-riding or ignoring consent. Under the reciprocity argument, I should not allow my data to be used, because I still benefit by free-riding on someone else's goodwill; but if everyone thinks that way, the social contract breaks down and no research will be carried out. Hence, the intervention of the state is not just desirable but is likely to be necessary to prevent public health research collapsing[59].

Within the public health profession there is therefore a broad consensus: in principle, public interest should be allowed to take precedence over consent. Note that this does not say what should happen in a particular case; the key issue is that consent should not be seen as necessary.

As it stands this is not controversial; even strong advocates of consent accept that this might not be the best outcome in all cases. Where the public health profession seems to differ is that this principle is strongly tied to the statistical concerns about using only consent. If consent is not necessary but is statistically problematic, perhaps the best solution is to devise the most statistically robust outcome and then see what needs to be done? However, this can be seen as a return to the 'paternalist' argument, and so most authors accept that reviewing the balance of public and private costs will remain an essential part of research approval.

It has been argued that considering the ethics of individual applications is missing a substantial trick; what matters is the overall risk to the public[60]. The introduction of the Western Australia Data Linkage System (WADLS) led to a large increase in the number of projects using (potentially identifiable) linked health data. However, it led to a large fall in the number of name-identified research projects, as researchers were able to use the pseudonymised data from the WADLS. It would be hard to argue that the overall public good has not been served by replacing a small number of projects having access to named data with a much larger number of projects only having access to match codes. The Menzies Foundation consultation on data linkage makes a similar point[61].

Finally, it should be noted that for linked data, much of the literature concentrates on the use of administrative data. Unlike statistical data collection, administrative data is the by-product of the primary purpose of serving the customer. Hence, for example, a GP may consider that doctor-patient confidentiality is his or her primary responsibility, not supporting the health service's research programme.

A3.3.3 Cultural barriers

This section considers a number of ways that the organisation and attitudes of data owners, researchers and the public affect the success of data linkage.

[59] Economists would recognise this as a standard analysis of a 'public good', in much the same way as other communal goods such as defence or police services; the need for intervention is a standard (and uncontroversial) result

[60] Trutwein B. and Rosman D. (2006) "Health data linkage conserves privacy in a research-rich environment". *J Ann. Epidemology*. V16:4

[61] https://www.chf.org.au/sub-1120-Menzies-Foundation-Consultation-on-Data-Linkage-Oct-2013.chf

Public attitudes to data sharing

Irrespective of the legality of data sharing and the ethical considerations discussed by RECs, public expectations can have a profound effect on the prospects for data linkage. For example, in the UK in 2014 a plan called "care.data" was unveiled to improve the use of GP data for research. This project became a source of much media interest; whilst scientific journalists made careful critiques of the plans, in the popular press boiled down to the question "can the government take your GP data and give it to whoever it likes, including private and insurance companies?" As a result of public concern, the programme was effectively put on hold with little serious public discussion over the programme's pros and cons, or whether the shortcomings could be addressed[62]. More importantly, public health professionals reported that the extremely negative reaction has made data owners more wary of data linkage generally[63].

Linking datasets can be more problematic in the public's eye because it immediately brings to mind the image of a government actively trying to find out more than the individual is prepared to disclose. In theory, gaining consent rather than using research gateways in legislation can legitimise the linkage in the public eye. However, as was noted above, consent is a human reaction to circumstances. One of the elements of care.data emphasised the 'opt-out' nature of the data plan, so that patients would need to inform their doctor that their data could not be re-used and linked. As part of the fallout of care.data, the implicit assumption of opt-out schemes (that patient data is a public asset unless the patient objects) came under significant scrutiny and the 'norm' of opt-in schemes was reinforced.

Public attitudes to data sharing can also be affected by non-research matters. The Walport-Thomas Review of data sharing[64] found that the UK public had low expectations of government's ability to handle personal data securely. This review came out shortly after the UK tax department had lost two CDs containing names, addresses and bank details of some ten million households (although the Review also noted that the low public expectations appeared to be long-established); this clearly affected the public attitude to data security, although the loss was due to failings in the tax department's administrative systems rather than the research use of data.

Public attitudes to sharing health data are regularly studied, both by health organisations and by academics[65]. The results are very sensitive to the specific situation. For example, questions about the value of security measures seem to get different responses in the US and Europe, even if worded the same way; but this seems to be because the primary 'human right' being violated is seen as 'privacy' in the EU, whereas US citizens are more focused on 'freedom of expression'[66].

---

[62] See Torjesen I. (2014) "NHS England postpones roll-out of care.data programme by six months", *British Medical Journal,* February, and related articles

[63] For example, O' Dowd A., (2014) "Patients could withhold information from GPs because of confusion over care.data scheme, doctors warn", *British Medical Journal,* February.

[64] Thomas R. And Walport M. (2008) *Data Sharing Review Report*. Health and Social Care Information Centre. http://systems.hscic.gov.uk/infogov/links/datasharingreview.pdf/view

[65] For example, Wellcome Trust (2013) *Summary Report of Qualitative Research into Public Attitudes to Personal Data and Linking Personal Data*. July. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_grants/documents/web_document/wtp053205.pdf ; Coppen R. (2015) "data protection and improvement in health through researcher: citizens' views". Presentation to CPDP 2015. https://www.youtube.com/watch?v=tRmODJ5lmDw

[66] Sarah St Vincent, (2014), Centre for Democracy and Technology, speaking at CPDP 2015 (not available on line)

There are some general conclusions that can be drawn[67]: the public is broadly:

- comfortable with health data being shared within health organisations;
- slightly less comfortable with research use;
- less comfortable still with commercial use;
- concerned about the security of their data;
- unable to distinguish between operational and statistical use;
- unable to distinguish between levels of anonymisation;
- much more positive towards health research than other uses of personal data;
- happy to change its mind (usually more pro-sharing) when provided with more information.

However, one factor dominates the public's attitude to data sharing: trust in the institution holding or sharing the data[68]. Gaining consent is also strongly positively associated with support for data sharing, but this is broad consent, not narrow; in other words, people are 'trusting' the organisation to do the right thing. Health providers (at least in the public sector) are usually at the top of the list of 'trusted organisations'[69].

Two large European projects, PriSMs[70] and SurPRISE[71], carried out surveys in multiple countries over some time, and come to much the same conclusion: if people have trust in the institution, then they tend to be very comfortable about decisions taken on their behalf. Of more relevance for the purposes of this report, health organisations come across again as highly trusted.

The disparity between general concerns about data sharing and support for specific ideas is partly a reflection of the more widely observed phenomenon: people tend to look more favourably on things they have personal experience of, and rely upon media reports for more abstract concepts[72]. More importantly, the answers depend upon both the framing of questions and the cultural background.

Risk aversion amongst data collectors

Making research use of confidential data involves risk. In general, data collectors tend to approach risk more conservatively than do the research community; this reflects the potential gains to each from research, and the potential losses to each from a breach of confidentiality. Some of these differences arise from differences in knowledge about the cost of a breach or the benefit of research, but there is also evidence to suggest that employees in public administration, who are often holders of socioeconomic data, tend to be more cautious in their outlook than other

[67] For a UK-centred but wide-ranging survey, see GMC (2007) "Public and Professional attitudes to privacy of healthcare data: A Survey of the Literature", carried out on behalf of the UK General Medical Council. http://www.gmc-uk.org/GMC_Privacy_Attitudes_Final_Report_with_Addendum.pdf_34090707.pdf

[68] For example, Brown Trinidad S, Fullerton S., Bares J., Jarvik G., Larson E. & Burke W (2010) "Genomic research and wide data sharing: Views of prospective participants" *Genetics in Medicine* (2010) 12, 486–495

[69] For example, Ipsos Mori (2014) "Public attitudes to the use and sharing of data" Report summary for RSS, June. https://www.ipsos-mori.com/researchpublications/researcharchive/3422/New-research-finds-data-trust-deficit-with-lessons-for-policymakers.aspx ; GMC (2007), ibid.

[70] http://prismsproject.eu/

[71] http://surprise-project.eu/research/

[72] Haddow G., Bruce A., Sathanandam S. & Wyatt J. (2011) "'Nothing is really safe': a focus group study on the processes of anonymizing and sharing of health data for research purposes" Journal of Evaluation in Clinical Practice, v17:6, pages 1140–1146

employees; and this, along with the lack of clear benefits, is manifested in strong resistance to data access[73].

It could be argued that in public health both parties are well aware of the *value* of research. However, this does not mean that *interests* are aligned. Consider the case of requesting co-operation from GPs to gain access to practice data. The GP might be aware of, and support, the idea that research on linked data brings substantial health benefits. On the other hand, if something goes wrong with the data, he or she would be accountable to the patients. It is individually rational to support public health research whilst refusing to commit oneself to providing the data. This model of 'diffuse benefits, specific costs' is held as one of the reasons why public servants appear to act in more risk-averse ways than others.

Academic perspectives on confidentiality

The perspectives of data owners on confidentiality are heavily influenced by half a century of academic studies into the disclosure risk associated with the release of data. Almost all of this literature uses 'intruder' scenarios: that is, it is postulated that a statistical expert with malicious intent will attack a statistical output in the hope of uncovering confidential information. In the simplest scenario, the intruder's only purpose is to embarrass the data owner.

This is a sensible approach to take when discussing the merits of different statistical techniques; it provides a common base against which alternative strategies can be compared. However, it has no empirical support. A very small number of publicly released tabular outputs have been used (or misused) to identify individuals; there have also been cases of individuals abusing their access to administrative data; and there are examples of researchers making mistakes or ignoring procedures to make life easier for themselves; but there have been no attempts to breach confidentiality by authorised researchers granted access to data for statistical purposes.

The lack of empirical support would not matter if the academic research was only taken as a guideline, or an extreme scenario. Unfortunately, this is not the case. The intruder model is popular with data owners, because it provides a worst-case scenario; if data is being managed securely even in the worst case, then surely the data owners have done their duty? This argument has merit if only the risk of breach of confidentiality is considered, but it clearly does not seek to balance public benefit against confidentiality protection[74].

Disciplinary differences

Medical data are analysed by public health specialists; socioeconomic data by social scientists; geographical data by geographers. Whilst there is pressure for researchers to be interdisciplinary (and such research seems to be increasing, although it is not clear if there are any formal measures), the standard working practices of disciplines encourage working with others in the same field. Meetings tend to be single-discipline, apart from cross-cutting technical events such as the big statistical society conferences.

---

[73] Ritchie F. and Welpton R. (2012) "Sharing risks, sharing benefits: Data as a public good", *Work session on statistical data confidentiality 2011*
http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/presentations/21_Ritchie-Welpton.pdf

[74] For an extended critique of the academic perspective see Hafner H-P, Lenz R. and Ritchie F. (2014) *User-centred risk assessment* Paper for Conference of European Statistics Stakeholders, November.

Moreover, it is hard for researcher to begin working cross-discipline from scratch. They need to identify co-workers in other disciplines before they can begin to review whether there is the scope for cross-disciplinary research, which can create a chicken-and-egg situation.

Data linkage can provide a spur to cross-disciplinary working because the ability to exploit data from different disciplines could encourage collaboration. Again, however, there is the chicken-and-egg: how does one know what data to link to generate a fruitful cross-disciplinary collaboration?

A3.4 Aspects of data linking: summary

Linked data suffers many of the same problems as single-source data: how to persuade data owners to release data, how to identify the correct legal framework, how to store the data and provide access to users, how to clean the data and spot errors, how to encourage effective use. On these topics, linked data is often more complicated in degree than in principle.

However, the fact of linking data from different sources does bring a number of additional problems to bear. The main one is the co-ordination problem: trying to convince two or more data suppliers to jointly concede control and placing their data into a more sensitive dataset. A less obvious problem is the nuances of Big Brotherhood that linked datasets bring to any discussion, particularly in the public arena.

In terms of statistical theory, the main issues of data linking seem to have been solved. While there are researchers looking into new linking methods, for all practical purposes the interesting questions were solved many years ago. Even in the area of consent there is little controversy amongst public health researchers: it is easy to find cases where insisting upon consent destroys the statistical foundation of analysis.

Operational problems do remain; in particular, identifying whether there is any selection effect in the linked data (irrespective of whether consent is a factor), and if so whether it is likely to affect outcomes. Practical problems such as cleaning data clearly exist but generally are seen as specific problems to be dealt with via user guides or other documentation for researchers.

Where there are still large unanswered questions is in the institutional framework. A consensus that informed and specific consent is not strictly necessary in all circumstance has not translated into a consensus on whether it should still be taken as the default, or whether public or statistical considerations should be the starting point. Academic research suggests that citizens are reasonably comfortable with research carried out by trusted institutions; but those institutions themselves are not necessarily comfortable with releasing data, particularly if they are not part of the public health community.

Finally, the summary of literature above is dominated by the news and research from high-income countries. These findings do not necessarily translate to low- and middle-income countries. Part of the aim of this project was to identify whether there were lessons that could be transferred between countries with different cultures, economics and models of governance. These are considered it the next section

Annex B: Data collection strategy

B1. Literature search

For the initial literature search on health and data-linkage was conducted, articles were obtained and analysed for relevance and suitability for the project based on the following search terms

- "Data linkage"
- "Medical Record Linkage"
- "Record Linkage"
- "Public Health"
- "Environment"
- "Preventive Medicine"
- Institution*
- Barrier*

The search terms were generated through controlled vocabulary check through the US National Library of Medicine's vocabulary thesaurus (NIH US National Library of Medicine, 2014). The search terms were also checked with experts in the field of data linkage and health to ensure they were comprehensive.

B1.1 Sources for the collection of data

The literature was obtained through online consultation of the references from the following bibliographical databases: Assia, Cochrane, Web of Science, Econlit, Medline (via EBSCO). These databases were chosen due to their relevance to health and/or data linkage research. A starting time was specified (published within the past 10 years) records were searched up until the 11.11.2014.

B1.2 Study Selection

Papers were screened by reading the titles and abstracts and removing all irrelevant studies using the inclusion criteria outlined below. The remaining papers were then read, and further separated into associated data linkage projects or countries. This then informed the researchers which countries/ projects appeared to be successful in supporting and facilitating published research, and which countries/ projects linked databases external to medicine.

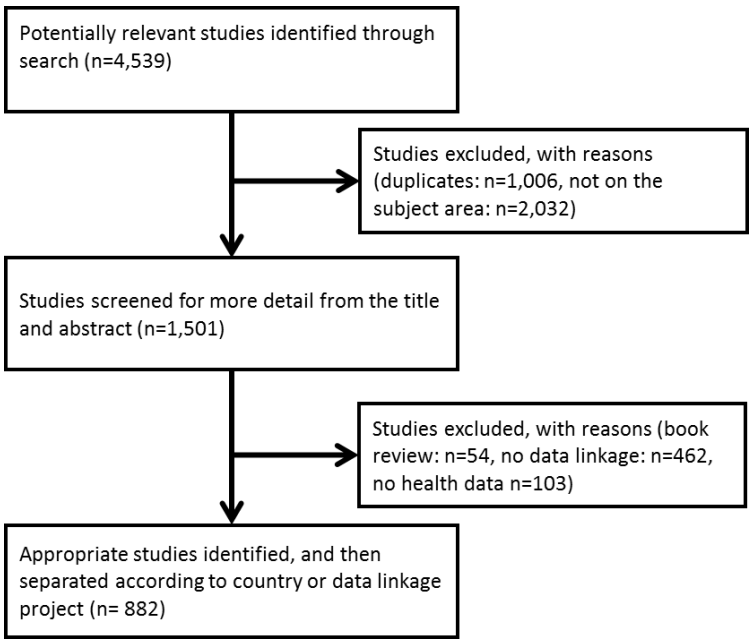
B1.3 Inclusion criteria

The inclusion criteria were identified and outlined through discussions within the research team. Studies must involve linking two or more separate data sets, one of the data sets must be related to health (e.g. data generated through health services). Since journals can demonstrate bias by only publishing significant findings, unpublished/ grey papers were included within this search. The study must also demonstrate evidence of impact- such as where there has been a public health outcome or health service design change.

B1.4 Results

Figure 4 below depicts the results of the search.

Figure 4 Flow diagram of systematic selection of articles



The search identified 4,539 studies, of which 1,006 were duplicates and 2,032 were not relevant to the subject area, these studies were excluded. Of the remaining 1,501 articles 54 were book reviews, 462 did not conduct/ contain data linkage, and 103 did not contain any health data therefore these were excluded. The remaining 882 papers were separated in categories according to country or data linkage project.

From the findings of the systematic search, location and project trends within the results became apparent. The search results showed that the Data Linkage Western Australia, data linkage service supported the most articles (n=182), followed by the US which supported 99 articles. Through identifying countries and projects which appear to facilitate data linkage this generated interest within the research team, into identifying the causation for successes within these research projects and how they counter-acted barriers to research facilitation and engagement. It is also generated interest to understand the reason why data linkage is not currently being practised within particular countries (e.g. India- which presently has no published data-linkage articles).

A theme which was prevalent within the search results was no problems and issues surrounding data linkage were reported within the published articles abstract or titles. This was contradictory to informal conversations with some authors from the articles (within the final result), cited a plethora of different issues including (but not limited to): institutional barriers, sustainability issues, administrative issues affecting their ability to obtain/ perform data linkage. Although the formal published articles failed to generate substantial barriers and facilitators to data linkage, the grey literature (unpublished reports and presentations) did provide insight into specific project barriers.

Note that that this may partly be a function of the review strategy. Because so many papers were generated, the initial sifting of papers was done by looking at the abstracts only. Systematic reviews carried out by the some of the authors on a similar topic have shown that a detailed reading of papers is necessary to identify practical difficulties, possibly because these are seen by researchers as side-issues. This will be investigated more before the final report, but it is worth noting that more

reference to difficulties encountered in project summaries may be a way to develop shared knowledge on potential problems.

The majority of the search results linked two medical databases. Although this is data linkage, linking two medical databases may be more straightforward than linking one medical database and one administrative database. Linking a medical database with an administrative database can pose a host of potential issues and barriers which can prevent successful linkages (such as the necessity to keep data anonymous yet retain specific details about the individual such as age). Therefore when considering case studies, it was deemed more useful to focus on projects which have linked medical databases with administrative/ government databases (or medical databases across different organisations).

B2. Interview strategy

Semi-structured interviews provide the opportunity for in-depth responses within a structured framework. The interview schedule was developed to prompt the interviewer to cover topics, surrounding the research aims of the project which included and were not limited to; experiences of data linkage, barriers encountered with data linkage, facilitators of data linkage. The interview schedule was peer-reviewed by academics prior to a pilot study involving a researcher using data linkage; the interview was then reviewed and deemed adequate in prompting response

Both snowball and opportunistic sampling was utilised to obtain participants. Members of the Public Health Research Data Forum received an email outlining the project and inviting them to contact the research team and subsequently arrange to be interviewed. Upon receiving the emails, some researchers consented to be interviewed and also provided further contacts within data linkage who may wish to be interviewed.

When reporting the data a different stance has been adapted from the normal protocol of addressing the respondent as a 'principal investigator based at x-organisation'. In this report no direct quotes are used unless the participant has agreed to be identified, as in the Case Studies. The reason for doing this is because the participants are potentially recognisable by the comments they have made and the description of the work they do. Data linkage is a specialist community and participants' descriptions may mean others in their field of work will be able to identify them (known as 'deductive disclosure').

A simplified interview schedule is given below. A modified schedule was used for ethics committee members.

Interviewee, organisation, date:	
Interviewer(s)	
Can you describe to me your experiences within data linkage?	
Can you describe to me the advantages of data linkage for your area of public health research?	
Can you describe to me barriers you have encountered surrounding data linkage? Please consider • Legal	

• Ethical • Statistical • Operational • Institutional barriers In each case • If the barriers were surmounted, how? • If the barriers were not surmounted, could you work around them?	
Have you observed any significant changes in practice over time?	
Do you apprehend any other barriers that you are yet to encounter within data linkage?	
Can you suggest ways in which data linkage can be better facilitated?	
Is there anything else you would like to comment on?	

B3. Interviewees

We are grateful to the following interviewees, who spent between half an hour and three hours providing the team with their insights into various aspects of data linkage. Some of these have led to the case studies. These were reviewed by the interviewees for accuracy before inclusion, and the interviewees agreed to be identified as such.

The views of these individuals influenced the report in many ways, and have sometimes been referenced directly; however, because each is a specialist in his or her area, and so easily identifiable, no comments in the report are sourced to individuals. This was to done to allow individuals to speak freely, which they did.

A number of other data specialists were interviewed informally by telephone and in person, including members of the Public Health Research Data Forum. Time did not allow formal interview, but we are grateful for their insights.

This report is based on the authors' interpretations of the opinions of interviewees, and no particular opinion should be ascribed to any individual or organisation.

UK

- Andrew Boyd, the Data Linkage & Information Security Manager for ALSPAC
- Professor Chris Dibben, Director of the Longitudinal Studies Centre Scotland
- Daniel Thayer, Research Analyst at SAIL
- Dr. Beth Thompson, Policy Advisor, Wellcome Trust
- Dr. Julie Woodley, Senior Lecturer in Radiography, Chair of HAS Faculty Research Ethics Committee, University of the West of England
- Professor Basia Zaba, Professor of Medical Demography, London School of Hygiene and Tropical Medicine

Sweden

- Dr. Anna Joud, Post-Doctoral Researcher, Division of Occupational and Environmental Medicine, Lund university

- Dr. Martin Persson, Senior Research Fellow, Centre for Appearance Research, University of the West of England

Australia

- Professor Lin Fritschi, Department of Epidemiology and Biostatistics, Curtin University
- Professor Louisa Jorm, Director of Centre for Big Data Research in Health, University of New South Wales, Australia

South Africa

- Andrew Boulle, Associate Professor in the School of Public Health & Family Medicine, Public Health Specialist for the Western Cape Department of Health
- Mark Collinson, Senior Researcher: MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, University of the Witwatersrand

India

- Dr Sanjay Juvekar, current leader of Vadu HDSS in India.

Bangladesh

- Subrata K. Bhadra, Senior Research Associate, National Institute of Population Research and Training
- AKM Ashraful Haque, Deputy Director, Bangladesh Bureau of Statistics
- Professor Dr. AKM Fazlur Rahman, Executive Director, Centre for Injury Prevention and Research, Bangladesh
- Prof. Dr. Mohamad Mahbub Alam Talukder, Faculty of Accident Research Institute

Annex C: case studies

The case studies below have been checked by the sources for the accuracy of the information, and the interpretations put on interview responses by the project team. The sources were interviewed in their own right as personal experts in the field, and comments should not be taken as representing the views of any group or organisation.

C1. ALSPAC (The Avon Longitudinal Study of Parents and Children)[75]

C1.1 What is the base situation?

The Avon Longitudinal Study of Parents and Children (ALSPAC - also known as Children of the 90s) is a large birth cohort study established in 1991 based in Bristol, England. ALSPAC have followed-up the health, well-being and development of multiple generations of study family members. Follow-up has been intensive and broad, with information collected on everyday exposure characteristics (including diet, lifestyle, socioeconomic status, parent–child contact, and GIS data), health and social outcomes (including health, educational, employment outcomes) as well as the compiling of a large bio-bank (including samples of urine, blood and genetic and epigenetic data). The ALSPAC website contains details of all the data that is available through a fully searchable data dictionary[76]. ALSPAC are augmenting this resource through linkage to a range of routine administrative records. A recent focus of this work, is the 'Project to Enhance ALSPAC through Record Linkage' (PEARL) which aims to develop generalizable methods to link to and utilise routine health and administrative databases in observational studies, and to increase the understanding of the secondary use of routine data using ALSPAC as an exemplar cohort. Since ALSPAC began in 1991 it has generated a substantial amount of research publications and projects.

C1.2 What data linking has been done?

To date ALSPAC has linked participant data with vital life events data sets such as obstetric and birth records, ONS mortality data and cancer registrations. ALSPAC has also linked participant data with clinical data sources including: clinical Hospital Episode Statistics (HES), Clinical Practice Research Datalink, NHS primary and secondary care records, education records, and demographic GIS data. Furthermore, ALSPAC is currently working on plans to link to tax and benefits records held by the Department for Work and Pensions and HM Revenue and Customs, and criminal records accessed via the Ministry of Justice. Linkage is conducted by the data owners (e.g. the NHS Health and Social Care Information Centre); typically using deterministic processes utilizing a range of personal identifiers (date of birth, gender, postcode, NHS ID number), although this differs by data owner. ALSPAC also conduct bespoke linkage projects, led by the ALSPAC Data Linkage Team, to locally held records using probabilistic and anonymized linkage methods.

C1.3 What were the factors associated with success?

C1.3.1 Institutional factors

A factor which was associated with success was the ability to demonstrate that ALSPAC is a reliable custodian of data. ALSPAC demonstrate this through 'Safe-Haven' governance structure and through certification to recognized information security standards (ALSPAC are certified to both the ISO 27001 and NHS Information Governance Toolkit standards). Being a recognized data holder encouraged organizations to share data with ALSPAC. Additional reassurances were gained through ALSPAC's engagement and transparency with study participants; enabling the study to demonstrate

[75] For further information see http://www.bristol.ac.uk/alspac/

[76] For further information see http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/

high levels of participant support for data collection via linkage methodologies. Local research ethics committees and the Health Research Authority Confidentiality Advisory Group have supported ALSPAC's linkage activities through providing ethical approval for linkage protocols and recommending the study receive 'Section 251' support to access the health records of participants who do not explicitly respond to requests for consent.

Access to centralized records (such as HES) is well established, however primary care records are still held locally and while ALSPAC have successfully extracted primary care records at scale, the process was arduous and resource intensive.

ALSPAC is continually reviewing and exploring how to further develop its infrastructure and governance framework, and has received financial support in developing and integrating new systems developed through PEARL and other projects.

C1.4 Barriers to data linkage and how they were overcome:

C1.4.1 Institutional factors

Data holders are becoming more hesitant to share data, with continuing uncertainty about how to meet the requirements of the Data Protection Act. Government departments are becoming more demanding on the evidence required to obtain data. ALSPAC are informing participants of how the study intends to use their records, and offering a means to object. This process is complicated by continuing uncertainties of the merits of opt-in vs. opt-out consent models and their application to the requirements of the Data Protection Act.

Further advantages lay in that the study aims and data usage can be demonstrated as being in line with efforts to improve the public good. That data sharing is being requested from a study perspective, rather than a generic perspective, seems advantageous.

Access to HES and education records is well established as these records are centralized, but access to other records is hindered by internal uncertainties regarding data sharing, or is only available via specific localized solutions (e.g. micro-data laboratories) which don't allow for the linkage between records from multiple sources. Others (e.g. primary care records) are held across many data owners which are seeking agreements at a local level. Such individual institutional contracts take time to establish. Although once achieved, organizations show a good awareness of their data sharing contract with ALSPAC - ensuring clarity and consistency in sharing.

C1.4.2 Operational factors

ALSPAC's primary experience through PEARL is that barriers to linkage are regulatory rather than technical or operational. However, new statistical and infrastructural developments (see below) can be used to address regulatory concerns. Technological advances have resulted in new possibilities for linkage and analysis (such as the potential for primary care data to be centralized through practice software providers); however these advances also pose potential problems (seen in the implementation of 'care.data').

C1.4.3 Statistical factors

There have been proposals for 'K anonymity' (ensuring that privacy is preserved as any single individual record cannot be distinguished from at least *K* other individual records); this is not ideal as it is complex to implement and may degrade the utility of the data. ALSPAC are testing the use of

DataSHIELD, a distributed-data model which provides aggregated data to researchers through anonymous summary-statistics and provides a simple approach to analyzing pooled data. Statistical tools allow some complex analyses of individual-level data, while still keeping the data on separate machines within the data owners networks, thus allowing data owners to maintain control of the data.[77]

C1.5 What lessons should we take away from this?

1. The necessity of establishing lasting, project-level, relationships with organizations and explicitly stating terms and conditions of data sharing with the data custodians.
2. Demonstrating ability to be a secure and reliable custodian of data.
3. The need for clear and transparent communication with stakeholders (especially the public, and in this case study participants)
4. The need for continuing investment in data management support and researchers investigating improvements to infrastructure and governance frameworks.
5. The importance of perseverance in obtaining data.

C1.6 Information source

Andy Boyd, Data Linkage & Information Security Manager for ALSPAC

[77] Wolfson, M et al. (2010). DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *International journal of epidemiology*, dyq111.

C2. SAIL (Secure Anonymised Information Linkage), UK[78]

C2.1 What is the base situation?

The SAIL (Secure Anonymised Information Linkage) Databank was established in 2006, and receives funding from the Welsh Government's National Institute of Social Care and Health Research. The SAIL Databank uses multiple government datasets including; NHS Wales Informatics Service (NWIS), Public Health Wales, Primary Care GP Dataset, Welsh Demographic service.

C2.2 What data linking has there been?

SAIL has successfully uploaded and linked over 2 billion records from multiple health and social care service providers. SAIL uses exact matching through the individual's NHS ID number, with a probabilistic matching process used when a valid NHS number is not present; the actual matching is carried out by NWIS. Datasets are held in a repository on an ongoing basis, for use in many research projects.

In the last year, there has been a notable increase in data provision by primary care providers (PCPs), rising from 40% to 70%.

C2.3 What were the factors associated with success?

C2.3.1 Cultural factors

A crucial facilitator was gaining NHS support for the broad project aims as well as operational aspects. NHS Wales were extremely supportive of SAIL utilizing patient records for research. In turn, SAIL has helped NHS Wales utilize the benefits and cost-saving of data linkage. A positive relationship between the two bodies has been developed.

The increase in PCP participation appears to be the result of a successful drive by a dedicated GP engagement team, suggesting that attitudes to sharing data are amenable to change if appropriately structured information programmes can be designed and resourced.

C2.3.2 Operational factors

SAIL has an unusual ethics procedure model with an extremely fast approval rate compared to similar facilities. This has been achieved through the development of the Information Governance Review Panel (IGRP). When an organisation agrees to share data, they may choose to delegate the due diligence on the use of their data to the IGRP. When a researcher requires that data, approval is given directly by the IGRP on behalf of the original data producers. Hence requests for multiple datasets can be handled quickly by a single body with the delegated authority to make decisions. The IGRP consists of representatives from the British Medical Association (BMA), National Research Ethics Service (NRES), Public Health Wales, NHS Wales Informatics Service (NWIS), and a Consumer Panel.

The IGRP ensures that the data will be used only for public benefit. The panel also decides whether the proposed data to be used is appropriate for the proposed research and also if there is a potential risk for disclosure. Within the panel there can be a variety of opinions on what constitutes an

[78] For more information, see http://www.saildatabank.com/

appropriate use of data. This may delay decision-making, but the need to obtain consensus from a range of stakeholders is a strength of the process.

Some data providers also require that they will individually review each request to use their data. This is handled by SAIL on the applicants' behalf, so it doesn't require an additional application. It is typically a streamlined process, because the data providers know the IGRP is also diligently reviewing the application, so it rarely introduces significant delay to the process.

C2.4 Barriers to data linkage, and how they were overcome

SAIL staff have written a number of reports on the problems of collating data, including operational factors (such as incomplete datasets or variation of coding within the dataset) and institutional factors (such as organisational resistance or management)[79].

C2.4.1 Institutional factors

Non-medical organizations can be more reluctant to engage with data sharing and linkage, as they may not directly observe the benefits; in addition, the law on consent can generate confusion and misinterpretation. Major data providers are generally well-informed about the information governance requirements for anonymised research, but holders of smaller, local datasets may not be as aware of this. A strong engagement and communication strategy can help.

The "care.data" negative publicity did increase scrutiny and public awareness of databanks. It appeared to have some impact on data collection, and SAIL improved safeguards although there was no indication that the existing ones were inadequate. To overcome this, SAIL has had to review its communication policies to ensure sufficient communication with the public and transparency in the form of more internal and external audits.

Communicating the benefits of linked data research, as opposed to simply focusing on minimization of risks, is one of the most important factors in encouraging the sharing of data.

C2.4.2 Statistical factors

SAIL follows standard good practice by having a trusted third party (in this case, NWIS) carry out the data linking. However, this makes it difficult to assess the validity of linkages. While an initial validation exercise showed the linkage algorithm to perform well, it can be difficult to assess the results in a specific case, particularly when linking based on incomplete or poor quality identifiers. This is largely unsurmountable; while quality checks in specific cases could be carried out, in general this is one of the compromises made in 'trusted third party' models.

C2.4.3 Operational factors

Practically, data problems occurred as SAIL has no control over the source data. Occasionally data had been received by SAIL which had formatting problems including no ID or number guides, making it difficult to establish what different values mean. This is less of a problem with large databases from large organisations: these usually consist of higher quality data and metadata, presumably reflecting a large organisation's need for good data management. In contrast, smaller databases seem to have more error, perhaps due to a higher level of human error in data entry and construction.

[79] A full list of SAIL papers is at http://www.saildatabank.com/data-dictionary/publications, of which several describe the setting up of the SAIL systems and choice made.

A problem which is specific for Wales is that both English and Welsh addresses are used and interchanged within the datasets, therefore making probabilistic record matching difficult to achieve in certain cases. This is addressed in the algorithms, but it complicates coding compared to a single-language system.

C2.5 What lessons should we take away from this?

1. Designing an approval strategy which incorporates safeguards from multiple organisations into a single body can substantially reduce approval times while still providing appropriate accountability.
2. Collaboration with key data producers at a high level (e.g. NHS Wales) can substantially reduce the friction associated with acquiring data, although by itself it does not address all problems.
3. Attitudes to data sharing can be changed by appropriate communication, as seen in the improved participation rates from PCPs.
4. Communicating the benefits of linked data research is one of the most important factors in encouraging the sharing of data.

C2.6 Information source

Daniel Thayer, Research Analyst at SAIL and a member of all the major SAIL committees. His responsibilities include supporting research projects and providing guidance and training on using the SAIL system.

C3. Scottish Longitudinal Study (SLS), UK[80]

C3.1 What is the base situation?

The Scottish Longitudinal Study (SLS) contains data on a random sample of roughly 5% of the Scottish population, a quarter of a million people. Census, vital events and education data are maintained as a single databank accessible to researchers on a project-by-project basis; health data is added for specific projects on a time-limited basis. Researchers apply directly to the SLS for access and are approved by the Research Board, on which all the data contributors are represented. For most health data, exceptional procedures apply.

Data are accessible through a secure RDC located on Scottish Government (SG) premises, operating to common standards with other similar RDCs throughout the UK. The SLS also co-operates with the ONS Longitudinal Study and the Northern Ireland Longitudinal Study, two broadly similar services covering the rest of the UK but with differences in sample size and data scope.

C3.2 What data linking has there been?

The Census records (1991, 2001, 2011) for these individuals are linked longitudinally. This socio-economic data is then augmented by vital events data (births, deaths marriages), migration data, and education data. Census and vital events data are supplied by the General Register Office for Scotland (GROS), migration data from the NHS Central Register, and education data from SG.

In addition to the health data from the Census, additional information on health events from NHS records can be linked on a project basis. This data is not part of the 'core' dataset, and is subject to additional scrutiny by the NHS ethical approval panel (unlike 'core' applications which are directly approved by the SLS panel). The linking is carried out by the NHSCIC for the specific project.

Since the SLS began in 2008, there have been 67 research projects[81]. Most researchers are based in Scotland because of the need to access the secure facility in Edinburgh.

C3.3 What were the factors associated with success?

C3.3.1 Institutional factors

SLS worked closely with the Scottish Government and GROS to develop the facility, based on a model which had been operating successfully for several years (the ONS Longitudinal Study). The decision to extend the databank to educational data arose from the early involvement of the Scottish Government. The inclusion of all data depositors on the Research Board means that decisions on access can be taken at the Board, without the need for further referral except in cases involving additional health data. This potentially could make the Board unwieldy, but in fact has allowed a critical mass of knowledge to be constructed from the Board members, which new members can tap into. Approval by the Board typically takes about six weeks, which is on a par with international levels.

A key decision was to ensure that data depositors were involved at an early stage in the design of the system: approval processes, flows of data, management arrangements and so on. This meant

[80] Website: http://sls.lscs.ac.uk/
[81] See http://sls.lscs.ac.uk/projects/ for a list of current and completed projects

that the Research Board solely concentrated on the broad policy aspects of data access, rather than day-to-day operation.

When considering privacy and ethical issues, the Research Board takes into account previous processes that the application has gone through and the ethical scrutiny of the entire SLS design carried out when the project was setup. Thus, it is assumed that application has been passed through the Ethics Committees of other universities, and that those committees are competent to judge the ethics of the application. When allied to the pre-approval of the delivery mechanism (on-site access at GROS offices), this reduces the need for researchers to provide duplicate or redundant information on the SLS application form[82].

C3.3.2 Operational factors

Initial planning to separate out data paths for identifying variables and 'payload data', and the approval of those pathways by the depositors, means that the data updates are simplified, privacy is maximised and legal requirements are met through pre-agreement.

C3.4 Barriers to data linkage, and how they were overcome

C3.4.1 Statistical factors

The main barrier has been achieving an appropriate match with acceptable costs. SLS took the decision to impose strict criteria on the automatic matching quality, so that more potential matches fall into the 'uncertain' category and so are subject to clerical matching. This is felt to improve the match rate, but also substantially increases the cost (although it could be argued that, because the 'core' data are reused, this is a more appropriate position than one-off data linking).

C3.4.2 Operational factors

As with SAIL, SLS follows standard good practice by having a TTP (again, NHSCIC) carry out the data linking, but this means that SLS staff do not have direct contact with data sources and are unable to establish the validity of linkages. Linkage quality in medical data appears to be high, but the very small amount of identifying information on education records (age, gender, postcode) gives concern over higher (but unknown) likelihood of false positives and false negatives.

Because SLS are not in direct contact with the data sources, checking the quality of other variables can prove problematic; nor is there a direct mechanism to feed back data issues to the sources. Hence, scope to improve the quality of data collection is limited, although the team has responded by developing multiple imputation approaches for missing data. There is also less direct value to some data providers of this work: data provision is seen as 'goodwill' rather than a core responsibility.

Finally, the need to maintain privacy does have cost implications, with a higher staffing level than would be necessary for a data archive dealing only with pseudonymised research data. Maintaining a secure physical facility imposes costs on both the SLS and on the researchers who need to travel to use the facility. SLS has introduced a 'remote access' facility which allows users to develop syntax off-site and then have the SLS team run it for them; this has improved the user experience but increases service costs for the SLS. The team has also developed software and methods for creating

[82] Governance document: http://sls.lscs.ac.uk/wp-content/uploads/SLS-Governancev3.2_noheader.doc

synthetic data which the research can run on his or her own computer, allowing familiarity to develop[83].

C3.5 What lessons should we take away from this?

1. Having all the data depositors on the approvals board can simplify the approval process by giving that body the authority to take decisions.
2. Recognising work done by other bodies to approve projects avoids the duplication of application processes.
3. Designing data flows and processes from the outset, and getting those approved, means that day-to-day operations can be managed as an internal operational issue.
4. Precedent can be an important source of confidence about the security and value of widening data access.

C3.6 Information source

Chris Dibben, Director, Longitudinal Studies Centre Scotland, University of Edinburgh.

[83] http://www.lscs.ac.uk/projects/synthetic-data-estimation-for-uk-longitudinal-studies/

C4. Data linking, Sweden

C4.1 What is the base situation?

Sweden has a unique set of population and compulsory health registers including patient register (hospital and care records); cancer register; death register; and prescription register. In addition to these health registers there are over 100 national disease or condition- specific registries available (for example the national rheumatology register). Compulsory population registers within Sweden includes register of economic and socio variables and social insurance registers.

C4.2 What data linking has there been?

The Scandinavian countries (Sweden, Denmark and Norway) implemented the use of electronic medical records in the early 1990s, where patient level data is linked to the individual's personal identification number PIN (not social security number) allowing linkages to as many databases possible. Data linkage between registers in Sweden is greatly facilitated by the fact that identification numbers are given to every resident and are unique for each individual. These numbers allow deterministic record linkages between registries and subsequently Sweden has delivered a substantial about of research utilizing data linkage between the Swedish registries.

C4.3 What were the factors associated with success?

C4.3.1 Cultural factors

The general population are compliant with their data being used and this is reflected in a low opt-out rate observed by researchers. One study investigating the Swedish public's preferences for information and consent procedures found the majority of the participants (n=2,122) are willing to delegate some decisions to the research ethics committees[84]. This study further reinforces that the public are willing to delegate decision making towards research using health data to governing bodies.

Each county council is responsible for all health care within the region, and hence the collection and maintenance of their respective population healthcare data. Subsequently, the government provides periodic supplementary grants to councils for the purpose of supporting the delivery and governance of care. This has generated competition for individual counties to demonstrate the 'best' care provision, encouraging effective data management. [85]

C4.3.2 Operational factors

Swedish registries are extremely comprehensive and detailed. Swedish hospitals register on real time, operating on the same server allowing for national data co-ordination. If the data have a low potential for identification then an opt-out system is employed by researchers. This is achieved through the public advertising campaigns that provide information on how the individual can opt out

[84] Kettis-Lindblad, Å., Ring, L., Viberth, E., & Hansson, M. G. (2007). Perceptions of potential donors in the Swedish public towards information and consent procedures in relation to use of human tissue samples in biobanks: a population-based study. *Scandinavian journal of public health*, *35*(2), 148-156

[85] Fredriksson, M., Eldh, A. C., Vengberg, S., Dahlström, T., Halford, C., Wallin, L., & Winblad, U. (2014). Local politico-administrative perspectives on quality improvement based on national registry data in Sweden: a qualitative study using the Consolidated Framework for Implementation Research. Implementation science: IS, 9(1), 777.

of the research to prevent their data from being used. However, if the researcher is requesting data from identifiable groups (i.e. particular rare disease groups), then the ethics committee will request that informed consent be obtained from the individuals.

C4.3.3 Statistical factors

A Personal Identity Code (PID) makes linkage easy, at least in theory, as deterministic matching can be used.

C4.4 Barriers to data linkage, and how they were overcome

C4.4.1 Institutional factors

The private health services report only what is necessary, and it is not compulsory for them to provide data to the same extent as public providers. Subsequently if data is required from these sources, permission to data access is required from each organization and the researcher will be required to individually link the provided data. However private health providers in Sweden only constitute a small fraction of overall health services.

Although almost 100% of health records within Sweden are digitised there are several different IT systems that organisations may use. This can cause a lack of uniform information standards and classifications across various organisations. However, each county has its own solution to standardising the datasets from different platforms and such solutions can also differ between counties resulting in further compatibility and contrast difficulties.

Researchers may frame their research questions based on the ease of data availability rather than the suitability of population data for the research question.

C4.4.2 Statistical factors

Names of data variables and level of quality within the data can differ between counties and over time which can cause difficulty within linking data. It is essential that the researcher has experience within data and comprehends what they are linking.

C4.4.3 Operational factors

Linked variables are only retained for ten years from survey data – survey specific keys are created and there are no generated archives for this data.

C4.5 What lessons should we take away from this?

1. Public acceptance of researchers using their personal data facilitates the principles and practice of data linkage.
2. Research is encouraged by the successful practice of an opt-out policy and the public compliance in delegating data access to governing bodies.
3. The government data administration and management incentives for counties can help assure data quality; however there are challenges and differences in practice within and between the counties.

C4.6 Information source

Dr. Anna Joud is a Post-Doctoral Researcher at the department of Occupational and Environmental Medicine at Lund University she has worked extensively with the Swedish registers.

C5. Analysing Longitudinal Population-based HIV/AIDS data on Africa (ALPHA) Network[86]

C5.1 What is the base situation?

Within the low income countries there is a distinct lack of vital registration systems, and the majority of illness occur at home with no or limited access to health care; hence, illnesses may not be identified and cause of death verified. It is difficult to ascertain disease incidence and prevalence within developing world populations. Community based surveillance sites such as the ALPHA Network collect information about an individual's health and condition-specific status directly from the data subject; this is augmented by verbal autopsies to establish cause of death when there is no direct information already collected[87].

The ALPHA network aims to maximise the usefulness of data generated in community-based longitudinal HIV studies across sub-Saharan Africa. This project connects all ten current community-based HIV cohort study sites in Africa, located in Kenya, Malawi, South Africa, Tanzania, Uganda and Zimbabwe. The network delivers training to facilitate analyses of the demographic and health data collected by the sites, after harmonising to a common format, which also facilitates comparative and meta-analyses of the pooled data.

C5.2 What data linkage has there been?

Several ALPHA Network[88] sites have experimented with different approaches to data linkage between the demographic surveillance data and the data from medical facilities serving the study areas. Two of the sites also run their own HIV care and anti-retroviral treatment (ART) clinics and collect medical data through these onsite medical facilities. In these cases excellent linkage is achieved through deterministic matching of the same unique individual identifiers which are used in the facility medical data and the demographic and HIV status data collected in the household surveys. Network sites in South Africa have also accomplished high quality linkage (surveillance data to medical data) by deterministic matching of the national identity numbers.

However, not every ALPHA Network site has its own medical facility or is in a country which possesses national identity numbers. Several study sites have experimented with probabilistic linking of common identifiers used in routine health service records (name, village of residence, year of birth) with the equivalent fields recorded in demographic surveillance. But where it has been possible to evaluate the results against a "gold standard" sub-set, the results have suffered from low specificity (too many possible matches with similar probability scores). Therefore, some sites have been piloting "real time" record linkage between demographic surveillance data and medical records at the externally managed ART clinics that serve their sites. This approach was based on methods

[86] For more details see http://alpha.lshtm.ac.uk/

[87] Other community based projects differ in their data collection; for example, the related INDEPTH projects (see other case study) do not collect HIV status and so rely more upon verbal autopsies

[88] For details about linkage within ALPHA network see: Reniers, G., Slaymaker, E., Nakiyingi-Miiro, J., Nyamukapa, C., Crampin & others (2014). Mortality trends in the era of antiretroviral therapy: evidence from the Network for Analysing Longitudinal Population based HIV/AIDS data on Africa (ALPHA). *AIDS* (London, England), 28(4), S533.

pioneered in the Karonga study site in Malawi, where individual identifiers in the demographic surveillance system include the names of parents and siblings – this means that when an individual is encountered at a clinic, questions about names of these family members can be used to accurately pinpoint the individual's identity in the demographic data base at a later point in time. The other ALPHA sites are not as systematic in recording names of close family relatives as part of an individual's identification data, but in all cases it is possible to retrieve the names of co-resident members of an individual's household. The "real time" approach requires that a member of the demographic surveillance team is stationed at the clinic with a copy of the demographic data base on a portable electronic device, and after obtaining informed consent for record linkage, uses probabilistic matching of name and demographic details to identify a range of possible matches, and then narrows down the search by enquiring about co-resident household members. The "real time" method is promising in the ALPHA setting, but is computationally intensive and it is not obvious how it could be used outside of demographic surveillance studies.

C5.3 What were the factors associated with success?

C5.3.1 Cultural:

Each study site has an extensive community liaison team continuously engaging with the local population to explain study aims and needs in appropriate language; it also collects opinions from the communities. Hence there is continual two-way flow of information to engage the local population and generate positive co-operation.

C5.3.2 Operational:

A factor associated with successful linking was the ability to use deterministic matching, which was possible for the South African site (through the national identity number) and sites which also ran ART clinics as they used the assigned demographic surveillance data number as the individual's patient number thereby making of the two data sets possible.

Establishing good relationships with local ministries of health helped develop an understanding of the project aims and the research sites ability to safely store data; this in turn aided the projects' access to ART clinic data.

Delivering workshops across and within the sites helps develop the research community, and also provide data quality control by ensuring that the onsite researchers are confident in the analysis and interpretation of the data.

C5.4 Barriers to data linkage, and how they were overcome:

C5.4.1 Operational factors

Within the South African sites, linking refugee data was not as easy to achieve through deterministic matching because not all refugees had acquired a national identification number. Probabilistic matching on name, residence area and birth year was used in studies where there was no personal identifier available, but an important barrier for matching was the use of nick-names and false names by individuals attending HIV test centres and treatment clinics due to the social stigma attached to HIV. Low literacy rates, complexity of names and variation in spelling can further hinder

the rate of probabilistic matching. Residence data are not as useful in a rural African context, where there are no street names or post codes. Automatic matching by names and individual demographic variables was not a viable linkage method due to these problems. The 'real-time processing' discussed above was developed as a response to these problems and has proved its value, but it may not be a sustainable solution

Another operational challenge stems from the fact that the majority of health facilities in Africa, including ART clinics, do not have electronic record systems, relying mainly on clinic registers and log books. This means that inaccuracies in recording the information cannot be checked at the time that is the data are recorded, further potential errors may arise in interpreting handwritten records, and additional time and labour are required to digitise the raw data.

C5.4.2 Institutional factors

An institutional barrier encountered was gaining access and permission to run ART clinics by external providers as this required negotiation between the local ministries of health and the organization or charity providing the service. At a national level there can be hesitation to allow access from the perspective of ministries of health and national treatment programmes as analyses of the data may reveal gross inadequacies and portray the government healthcare system negatively. A further consideration is that charities and other organizations running the clinics may intend using the data collected themselves. In ALPHA's experience it is vital to take time to reassure the ministries that the data will be stored securely and to promote collaboration between the research teams and the health service organizations.

Allocation of research funds may also determine the possibilities of pursuing data linkage schemes – funding is generally distributed to local research institutions, but very often these institutions have overseas partners who are "in the driving seat" – either because they generate the research ideas, or because they handle the disbursement of funds, or both. But by definition data linkage projects involve sharing or obtaining data from entities that are part of quite different structures, such as health ministries or international agencies, and these bodies may also incur substantial costs in preparing data or adapting the protocols that they use in collecting data. These large institutions may have no tradition of receiving small amounts of research funding for demonstration projects, and may find it politically unacceptable to enter into negotiations with entities that may be perceived as junior partners of foreign universities. The benefits of data linkage projects at a national level must be obvious to make them attractive to the institutions whose data we want to link to, and this generally implies that there must be a vision of taking projects to scale, a willingness to explain the research data to the institutional partners, and a program for training staff from the counterpart organisation to use the linked data for their own purposes.

C5.5 What lessons should we take away from this?

1. Importance of local knowledge to enable positive ties with the local community and the authorities that hold the data that are to be linked.
2. Wider humanitarian concerns regarding stigmatised or vulnerable groups such as persons living with HIV or refugees and the additional difficulties faced in obtaining population data from these groups.
3. The need to (work hard to) ensure credibility as trusted partners to hold shared data.

4. The lack of reliable personal identifiers to act as record link fields impacts on prospects for extending the type of comparative analyses that have been the hallmark of successful collaboration between ALPHA study sites and their various clinical treatment partners.

C5.6 Information source

Professor Basia Zaba of the London School of Hygiene and Tropical Medicine is the founder and principal investigator of the ALPHA Network.

C6. Western Cape Department of Health, South Africa

C6.1 What is the base situation?

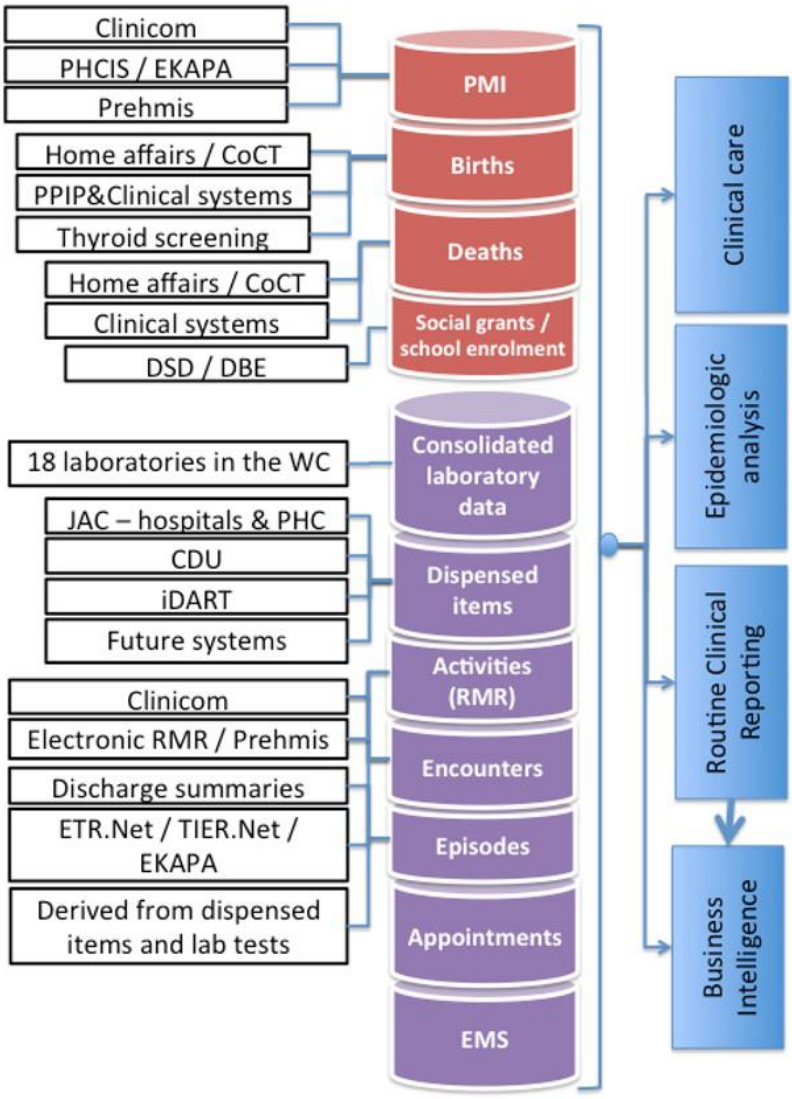
The Western Cape Department of Health (WCDoH) has developed an integrated health data system in the province. This project seeks to evolve the patient registration system through the creation of a shared folder number for each patient, which allows for individual level linkage across several different information systems (pharmacy, clinical records, civil registration, laboratory and disease registers). Previously, Western Cape health data had generally not been available at an individual/patient level and data was not easily linkable; the focus was on using aggregated data against cost centres to analyse the efficiency of the overall system. This change has been made possible as a result of the near pervasive implementation of a unique health identifier when patients register at any Provincial or City facility. This unique number is called the Patient Master Index (PMI) and the centralised data repository is simply called the “data centre".

C6.2 What data linking has there been?

Data from Western Cape hospital information systems that utilize the PMI (Clinicom[89], PHCIS[90]/EKAPA[91], Prehmis[92]) are available for linkage on the system. The WCDoH and the Department of Social Development (DSD) also aims to collaborate on access to birth registers and maternal ID numbers. The WCDoH has developed a Burden of Disease surveillance system which captures fact, date and cause of death from duplicated copies of death certificates. These data are in turn linked to the PMI, and are then available both for burden of disease estimation, and as outcome data for health program evaluations. However there is a proposal to link the database to social grant and school enrolment data. The system also captures data from clinical domains. Laboratory data are linkable to patients irrespective of whether the laboratory requests were from hospitals, provincial clinics, or City of Cape Town Clinics. Data on drug dispensing, “encounters" (e.g. clinic visits, hospital outpatient visits, hospital admissions, community care visits, and phlebotomy or dispensing visits), routine indicator data, emergency medical service, appointments, and episode data have all been incorporated into the system to varying degrees (or there are plans to incorporate those data). There are even plans to link radiological images in future. The figure below summarises these data domains and their sources.

[89] The hospital information system in the majority of hospitals in the Province
[90] Primary Health Care Information System, the system used for patient administration and routine information collection in Provincial primary care clinics
[91] The Provincial centralised/online solution for HIV and TB monitoring, on the same platform as PHCIS, and currently being merged with PHCIS, and further developed as a multi-disease monitoring platform and electonic medical record system. Name derives from the first version (Evaluation of the Khayelitsha AIDS Programme)
[92] Patient Registration and Health Management Information System - the system used for patient administration and routine information collection in City of Cape Town primary care clinics

Figure 5 Source systems, and demographic and clinical domains (Source: Western Cape Department of Health, 2013)



C6.3 What were the factors associated with success?

C6.3.1 Cultural

The project is premised on good collaboration between government and academia. Key figures in the project are involved in both the School of Public Health and Family Medicine at the University of Cape Town as well as operating as public health specialists with the WCDoH. Previous research work linking HIV cohort data from research databases to national population registers demonstrated the value and manageability of such projects[93].

[93] See:
Boulle, A., Cutsem, G. van, Hilderbrand, K., Cragg, C., Abrahams, M. & others (2010). "Seven-year Experience of a Primary Care Antiretroviral Treatment Programme in Khayelitsha, South Africa". *Aids*, 24(4), 563{572.
Cornell, M., Lessells, R., Fox, M., Garone, D., Giddy, J. & others (2014) "Mortality among Adults Transferred and Lost to Follow-up from Antiretroviral Therapy Programmes in South Africa: A Multicenter Cohort Study". 67(2).
Cutsem, G. Van, Ford, N., Hildebrand, K., Goemaere, E., Mathee, S. & others (2011). Correcting for Mortality Among Patients Lost to Follow Up on Antiretroviral Therapy in South Africa: A Cohort Analysis. *PLoS One*, 6(2).
Johnson, L., Mossong, J., Dorrington, R., Schomaker, M., Hoffman, C. & others (2013). Life Expectancies of south Africa Adults Starting Antiretroviral Treatment: collaborative Analysis of Cohort Studies. 10(April).

Linking ID numbers of identified individuals to South African national population registers, as was required in the above projects, often would have to go through the Medical Research Council (MRC). The MRC has a long-standing role in linking South African CR data. The experience with both public health data information systems usage and research was valuable in forming the necessary inter-institutional relationships that allowed those systems to be successfully linked as part of the data harmonisation project.

Before the creation of the data centre there weren't many attempts to link at the individual patient level. Only in the last three or four years has the coverage of the PMI become so good that the aspiration to link everything has surfaced.

C6.3.2 Statistical

Changes that allowed for the linking of individual level data across multiple source systems are[94]:

- a near pervasive implementation of a unique health identifier when patients register at any Provincial or City facility, resulting in electronic data collected at any of these facilities for a single patient being linkable, including laboratory tests and medicine dispensing ordered against this identifier;
- the availability of consolidated laboratory data from all 18 laboratories in the Province through the NHLS Corporate Data Warehouse (CDW);
- an appreciable proportion of medicine dispensing being completed through electronic systems, at hospitals, through the chronic dispensing unit (CDU), and more recently at primary care facilities;
- increasing proportions of patients with civil identification numbers recorded, which enables linkage with administrative systems from other departments, including births and deaths (Department of Home Affairs), social grants (Department of Social Development) and school enrolment (Department of Education);
- the expectation that primary care routine activity data collected electronically in future.

The introduction of PMIs is perhaps the most critical introduction. These PMIs are common across source electronic patient systems. When patients register for the first time at any health facility all systems are able to check the PMI through a wide area network (WAN) to see if the patient has already been captured. If the patient does not already exist on the system, they are issued with a fresh PMI.

Some patients are occasionally issued with more than one number erroneously but it is possible to check for these duplicates probabilistically using other variables (name, surname, national identification number and date of birth). There is a proposal to integrate more systems into the source data (including social grants, administrative data, school data), all of which would increase the amount of check data available and increase the rate at which duplicates are successfully identified. However, it is likely that business process changes offer more scope for improvement, for example by checking for PMI duplication at the entry stage, rather than the search stage.

[94] Western Cape Department of Health. 2013 (April). *Strategic Approach to Patient-Level Health Data Harmonisation and Integration*.

C6.3.3 Ethical

Ethical approval for the linkage was not cited as a huge constraint, as the linked data is primarily used for clinical care (which is approved internally by the WCDoH). The protection of personal information is a concern for the WCDoH in using the data to produce research, but less so when using the linked data for clinical purposes.

Currently, the WCDoH is registering the data centre with the UCT ethics committee for approval of data management, curation and protection processes. This will form the backbone for individual human research ethics committee applications of project specific research proposals, which would each have to consider issues of patient protection, risk and benefits of the research.

The ambition is to further establish a facility which could make anonymised data available to researchers with projects which have received appropriate ethical and governmental approval, operating across government departments. The WCDoH is currently collaborating with the Western Australian linkage project and the Farr Institute in the United Kingdom to realise this. The immediate goal is to link health, social services, education and population registers.

C6.4 Barriers to data linkage, and how they were overcome

C6.4.1 Statistical

The relatively wide coverage of national ID numbers (80 %) is useful for linking some data, but these ID numbers are not always available to health services. Deterministic matching on ID might be prone to selection issues given the probability of having an ID is correlated with a number of variables of interest. Probabilistic matching also has some problems, including:

- poor capture of date of birth;
- anglicised first names and African first names used in different data sources;
- married and maiden names used in different data sources;
- twins having similar first names (commonplace amongst certain South African cultural groups) and identical date of birth.

C6.4.2 Operational

High level data skills are almost completely absent from the DoH. Management of complex information systems will usually be outsourced to an organisation which has very little incentive to diligently and accurately match and link.

Pay is an issue for data professionals. Other cadres of professionals (e.g. doctors, engineers) receive an occupation-specific package which is delinked from the management hierarchy. This means it is possible to get a doctor being paid more than a director, despite the director being more senior in the organisation hierarchy. This de-linkage has not been operationalised for computer engineers, software engineers, data scientists, etc. The highest pay level the department can employ a SQL programmer or a software programmer is far less than they'd be paid in the private sector. The only way the government does employ technical people in the information sciences is on contract or by outsourcing via some sort of tender process. While the South African State Information Technology Agency (SITA) does contract in developers and programmers on behalf of government to try and service some of the cross-departmental functions, the transversal nature of the agency means that there aren't health specific data scientists being developed or nurtured as specialists. Ideally, one

would want to build up health information system experts who are embedded; sets of professional who, if they're not doing the work themselves, will have the skills to manage the appropriate outsourcing of the work.

C6.5 What lessons should we take away from this?

1. Positive relationships between organisations are important for joint projects
2. Successful ad hoc projects can provide the evidence base for the value of more strategic projects
3. It is worth spending time on improving the accuracy of the probabilistic link fields
4. A strategic approach to collecting match fields across organisations pays dividends, but…
5. Even in exact matching one needs to be aware of the potential for non-random missing match variables.

C6.6 Information Source

Andrew Boulle, Associate Professor in the School of Public Health & Family Medicine, Public Health Specialist for the Western Cape Department of Health

C7. The Agincourt Health and Socio-Demographic Surveillance System (HDSS), South Africa

C7.1 What is the base situation?

The Agincourt health and socio-demographic surveillance system (Agincourt HDSS) was established in 1992 and is located in rural North-East South Africa near the border with Mozambique. It provides the foundation for the Rural Public Health and Health Transitions Research Unit of the Medical Research Council (MRC) and University of the Witwatersrand, South Africa (the MRC/Wits-Agincourt Unit) who are also responsible for funding other health data linkage events. The Agincourt HDSS annually captures household roster information, pregnancy outcomes, mortality/deaths (by verbal autopsy), migration, maternity history and union status, as well as a variety of other social variables which are captured periodically (labour force participation, education, etc)[95].

C7.2 What data linking has there been?

Agincourt HDSS is regularly linked to other data sources. In collaboration with the South African Department of Home Affairs (DHA) and Statistics South Africa (Stats SA), data from the Agincourt HDSS was linked with national South African Civil Registration (CR) systems; this data linkage project required the signing of a non-disclosure agreement and for one of the senior statisticians to access a secure data centre at Stats SA in Pretoria. Agincourt HDSS frequently links its data with clinical data (from clinics within the enumeration area) stored in the provincial primary healthcare system. This requires collaboration with the South African Department of Health (DoH) and the clinics themselves. These clinics include primary health care units, HIV/AIDS treatment clinics, hypertension clinics and others. Linkages between Agincourt HDSS data and schools data have also been piloted within the enumeration area and required some collaboration with the Department of Basic Education (DBE).

C7.3 What were the factors associated with success?

Overall, Agincourt HDSS was in a good position to link otherwise difficult-to-access data with its data, but it is not clear whether other units or researchers would be able to replicate the unit's success.

A major factor underpinning the success of these various data linkages as the strength of the research unit's relationship with the various departments that act as gatekeepers to the data. The Agincourt HDSS project team maintain strong relationships with the DoH and the DHA, and have a generally high level of collaboration and mutual trust with the state: some team members have 20 years of experience working with government and the health research units. There is a substantial benefit from gaining the trust of government departments who may otherwise be wary of studies that could invite criticism. This institutional relationship was useful, for example, in getting approval from the DoH for the clinic record linkages as their ethics processes are internal to the department, and not amenable to external argument. The same is true of the Stats SA and the DHA project (the

[95] For more detail see Kabudula, C.W. and 12 others (2014) "Evaluation of Record Linkage between a Health and Demographic Surveillance System and National Civil Registration System in South Africa". *Population Health Metrics*; Kahn, K. and 18 others (2012) "Health and Demographic Surveillance System Profile, Profile: Agincourt Health and Socio-Demographic Surveillance System*". The International Journal of Epidemiology*.

CR linkage) which only required the Agincourt team to acquiesce to a set of conditions surrounding the usage of the data.

C7.4 Barriers to data linkage, and how they were overcome

Although all the attempts to link the Agincourt HDSS data have been successful to date, there are a number of potential barriers. Legal, ethical and institutional barriers did not seriously inhibit the success of the linking project, as the Agincourt team were trusted in the internal ethics appraisals of each of the departments. However, there were operational and statistical difficulties: in particular, skills shortages in information system administration and data capturing at the clinical level. This occasionally led to poor data capture or poor maintenance of servers. While it is difficult to be exact, the quality of data was potentially degraded (which naturally inhibits the success of probabilistic and even, potentially, deterministic matching if ID numbers are not correctly captured). There are also skills shortages in research when it comes to actually linking data itself.

Another factor that limited the success of their matching endeavours is the comparatively low rollout of South African national ID numbers in the North West. This is speculated to be the result of high refugee and illegal immigrant influx through the relatively porous Mozambique border. Non-nationals would, of course, not have ID books or ID numbers.

This introduces a potential for selection bias introduced by deterministically matching on ID numbers; the potential selection bias occurs because matches are correlated with individual characteristics that are of interest to researchers e.g. immigrants are less likely to have ID booklets and are therefore more likely to be excluded in the data matching process[96].

It could be worthwhile to introduce a private identification number system into health data information systems in South Africa, much like what has been proposed (and is being successfully implemented) in the Western Cape. The team, in collaboration with the DHA, did run an experiment trying to establish the areas where few people have ID booklets. This highlights both the relative difficulties of their data matching procedure but also importance of the level of trust between the AHSDS team and the DHA when it came to exploring solutions.

C7.5 What lessons should we take away from this?

1. Senior researchers' connections and experience in dealing with state departments was a crucial factor in establishing the projects for linking the data
2. There are significant benefits from closely interacting with government departments in data linkage projects
3. There would be value in a centralized body that improves access to potentially linkable data; if data producers could delegate the burden of necessary due diligence onto a trusted third party it could improve public health data linkage and the quality of public health research overall.
4. Such a body could introduce the added benefit of relieving skills-constrained research groups of the burden of having to actually link the data carefully and well.

[96] See also the complementary South African case study in this report

5. There would be significant gains from introducing a private identification number system into health data information systems in South Africa, much like what has been proposed (and is being successfully implemented) in the Western Cape.

C7.6 Information Source

Mark Collinson, Senior Researcher: MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, University of Witwatersrand; Co-Theme Leader: Households Responses to Shocks and Stresses. Co-Theme Leader: Demographic Levels, Trends and Transitions, Agincourt Unit

C8. INDEPTH Network, Africa/Asia/Oceania[97]

C8.1 What is the base situation?

The INDEPTH network is a network of health and demographic surveillance system (*HDSS*) field sites. The HDSS sites obtain information on all residences including updates vital events such as birth, death (through conducting verbal autopsies), migration, and other attributes such as relationship and age. The network presently have 54 centres across three different continents (Asia, Africa and Oceania) with some HDSS sites having clinical health facilities incorporated allowing for data to be captured on health service attendance, diagnosis, service and treatment. The INDEPTH member centres collate HDSS data and then share it within the public domain allowing approved researchers to access the data. Currently 21 of the centres place routinely collected core micro datasets into the iSHARE2 [98] which is INDEPTH Sharing and Access Repository; a web-based system allowing access the network data worldwide.

C8.2 What data linkage has there been?

The Agincourt HDSS branch of the INDEPTH network has conducted data linkage between national registries and HDSS data within Africa- for further details see the previous case study. This case study will focus on the sharing of the HDSS data and the factors surrounding data access at the HDSS sites that have considered or attempted data linkage of the HDSS data with national registries. Currently within the Vadu HDSS site there have been efforts to link and analyse the HDSS data with a Geographic Information System (GIS), and also linking of previous collected data sets with HDSS.

C8.3 What were the factors associated with success?

C8.3.1 Operational factors

Technology and software are developing quickly and institutions in LMICs may not have access to the appropriate software. Therefore by sharing the HDSS data through a public domain (the iSHARE2) other researchers with access to appropriate software can transform the data on behalf of the institution. The institution sharing data is also at an advantage as it means there are no resources spent in cleaning, analysing and publishing the data. Placing the data in the public domain through the iSHARE2 repository enables international collaboration with researchers and specialists across the world, thereby allowing for exploitation of the data. It also has the potential benefit to provide PhD students with data who may be unable to collect data, due to lacks of resources or living in a rural and remote area. Through sharing data it can allow an institution to showcase its ability within data collection, management and quality assurance skills

C8.4 Barriers to data linkage, and how they were overcome

C8.4.1 Cultural factors

Despite the benefits associated with sharing data there is still reluctance amongst some institutions to share data. Within the INDEPTH network it is hoped that all HDSS sites and members will be wholly sharing the collected data, however there has been scepticism as to whether this happens.

[97] For further details see: http://www.indepth-network.org/
[98] For further details see: http://www.indepth-ishare.org/

C8.4.2 Institutional factors

A factor which has proven to be a barrier in data access is that lack of awareness surrounding the value of data sharing and the potential benefits to be gained. It was felt that within lower income countries data linkage and sharing is not a concern of the ministries, despite both health and education ministries (within India) generating data. A reason for this is that the ministries do not view data sharing as an immediate priority due to other impending concerns. There were also reported administrative issues within macro organisations and uncertainty about the legalities of data sharing and confusion regarding how to physically share the data.

C8.4.3 Operational factors

However if funders want data to be shared they must fund researchers and allocate time after the project for researchers to clean and upload their data to the public domain. Currently projects are funded from start to finish with no consideration or funding for the time spent in cleaning and transforming data for the public domain.

C8.5 What lessons should we take away from this?

1. Funders need to allocate time and funding for researchers to share the dataset.
2. The necessity on advocating to data collectors and public health research organisations the benefits of sharing data.

C8.6 Information source

Dr Sanjay K Juvekar is an Anthropologist and current leader of Vadu HDSS in India. He conceived the concept of data sharing in INDEPTH network by initiating iSHARE repository and was first Principal Investigator of iSHARE supported by INDEPTH Network.

C9. CHeReL[99] (Centre for health record linkage) , Australia

C9.1 What is the base situation?

CHeRel is a data linkage research facility established in 2006 to create and maintain a record linkage system for health and human services in NSW (New South Wales) and the ACT (Australian Capital Territory) and is funded by the Population Health Research Network. This research facility is managed by the NSW ministry of health and is partnered with NSW Cancer Institute, NSW Health, ACT Health, University of New South Wales, University of Sydney and the University of Western Sydney.

C9.2 What data linking has there been?

CHeReL uses a master linkage key system (MLK) which routinely links data between numerous health data records within NSW and ACT; this includes hospital admissions, emergency department datasets and datasets containing information about incidence of diseases, conditions and routine health testing. The master key also links data surrounding vital events including birth and death records for NSW and ACT. External to the routine master linkage key system, CHeReL links data on an ad-hoc basis with study-specific databases, including the Australian Study of Women's Health which collects information on women's health, well-being and socioeconomic data. CHeReL holds over 93.9 million records based on 10.9 million people with an average 6.0 links per person.[100]

C9.3 What were the factors associated with success?

C9.3.1 Cultural factors

Demand from the research community coincided with advocacy from champions within key data custodian agencies, at a time when Australian funding for health and medical research was increasing, and the global financial crisis had not yet hit.

Australian research groups have published numerous validation studies[101,102] that explore the quality of linked data and provide guidance as to their appropriate use. This has helped to counter the prevailing view among grant assessors that administrative data are of "poor quality", and has increased researcher awareness of the processes by which administrative data are captured, and the rules which drive how diagnoses are coded.

There has been an observed increase in international interest and collaborations; for example, Scotland (ASH- Avoidable Scottish Hospitalisations) provides Australia with insight into the Scotland QOF (quality outcomes framework) through Scottish management of clinical performance data and Scottish hospitalisation records[103]. In contrast Australia holds detailed large-scale data from the 45 and Up Study cohort that can be linked with GP claims and hospital records, therefore allowing investigation of the roles of socio-demographic, lifestyle and geographic factors.

---

[99] For further information see: http://www.cherel.org.au/

[100] Centre for Health Record Linkage. (2011). *CHeReL- Master Linkage Key.* Available: http://www.cherel.org.au/master-linkage-key.

[101] Lujic S, Watson DE, Randall DA, Simpson J, Jorm LR. (2012). Variation in the recording of common health conditions in routine hospital data: study using linked survey and administrative data in New South Wales, Australia. *BMJ Open* 4: e005768. doi:10.1136/bmjopen-2014-005768

[102] Tran D, Jorm L, Lujic S, Bambrick H, Johnson M. (2012). Country of birth recording in Australian hospital morbidity data: accuracy and predictors. *Aust NZ J Public Health* 36:310-316.

[103] Jorm, L. R et al. (2012). Assessing Preventable Hospitalisation InDicators (APHID): protocol for a data-linkage study using cohort study and administrative data. *BMJ open*, 2(6).

C9.3.2 Operational factors

The CHeReL links records using probabilistic matching of the demographic details, and assigns a CHeReL personal ID number for records that belong to the same individual. The CHeReL personal ID and the associated source record numbers form the CHeReL MLK. Although there can be incidents of false positive links with MLK or incorrect information being provided from the source database, a 2012 review of the MLK[104] found that the overall percentage of individuals and records affected was substantially small (3/1000). The reason why CHeReL is successful is due to having a master linkage key which is continually updated with routine data allowing for enrichment of the data. When considering database management systems, the MLK system will theoretically incur less error in contrast to traditional methods where link probabilities are re-generated. Subsequently if a researcher is requesting to use data from the master linkage key then the process to obtain ethical and data custodian approvals and the data extraction process is reasonably fast.

C9.4 Barriers to data linkage, and how they were overcome

C9.4.1 Institutional factors

Data custodian agencies can be cautious and hesitant about data linkage and sharing, particularly where this involves data from more than one jurisdiction. This can be for various reasons - including uncertainty about the legalities surrounding providing data, and inconsistency in legal and policy frameworks among jurisdictions. In some Australian jurisdictions, enabling legislation is absent, or has not recently been updated, and is therefore "silent" about data linkage. The Australian National Health and Medical Research Council (NHMRC) is currently developing Principles for Accessing Publicly-Funded Data for Research[105] which will provide guidance and help aid data custodians' decision-making about data linkage and sharing.

Another institutional factor is the level of detail within the data that an organization is required to report; for example in NSW it is at the discretion of private hospitals as to whether name information is collected. Bentley et al.[106] reported missing name information for both mothers and infants which thus affected linkage rates. CHeReL generates linkage through probabilistic matching of demographic details (including name), and so improves linkage if organizations provides a whole data set.

C9.4.2 Operational factors

There can be an issue with non-reporting of variables when the coder is coding the patient's notes - an example is not coding the patient's ethnicity. This is not necessarily a data quality issue because if the variable has been previously reported then through the multilevel linkage the missing variable will automatically be replaced by the previous variable. However, CHeReL is updated through batch linkage - until the variable is added it will remain missing; it is therefore essential to emphasize good coding practice and ensure coders receive sufficient training.

[104] Centre for Health Record Linkage. (2012) Quality Assurance Report http://www.cherel.org.au/quality-assurance .
[105] National Health and Medical Research Council (2014) Draft Principles for accessing publicly funded data https://consultations.nhmrc.gov.au/files/consultations/drafts/draftprinciplesaccessingpubliclyfundeddata141209.pdf
[106] Bentley, J. P., Ford, J. B., Taylor, L. K., Irvine, K. A., & Roberts, C. L. (2012). Investigating linkage rates among probabilistically linked birth and hospitalization records. *BMC medical research methodology*, 12(1), 149.

C9.5 What lessons should we take away from this?

1. International collaboration can provide insight into areas not previously addressed.
2. Continually updating the MLK with routine data allows for faster access to data sets and extraction for researchers.
3. The MLK is most efficient (and hence better for researchers and subjects) when it is provided with all the detail to build good keys
4. Don't be frightened of administrative data: distinguish between errors in data (not common) and gaps in data (more common, but can be addressed)

C9.6 Information source

Professor Louisa Jorm an epidemiologist who played a key role in the establishment of the Centre for Health Record Linkage (CHeReL). She is also a member of the NHMRC Research Committee and also the Chair of Data Linkage Committee for the 10 to Men Study and the Chair of the Policy Advisory Group for Centre of Research Excellence in Women's Health in the 21st Century.

C10. The Bangladesh Demographic and Health Survey (BDHS), Bangladesh

C10.1 What is the base situation?

The National Institute of Population Research and Training (NIPORT) was established in 1978 with a vision to stand as a Regional Training and Research Institute on health, especially reproductive and child health in South Asia[107]. Its current mission is to provide task oriented in-service training to health & family planning program personnel and conduct program focused studies and operations research in Health & Population sector Program in Bangladesh. The overall goal of NIPORT is to contribute to improve the health status of families in Bangladesh, The purpose of NIPORT training and research activities is to make sure that program managers and service providers are effective and efficient in providing quality services on health, especially reproductive and child health care in the communities of Bangladesh. NIPORT's objectives are related to the overall goal of Health and Population Sector Program in Bangladesh, which is to improve the Health and Family Welfare by birth spacing and better status, particularly of mother and children.

The Bangladesh Demographic and Health Survey (BDHS), part of the worldwide Demographic and Health Surveys program (MEASURE DHS), has been running in Bangladesh since 1993. Currently, BDHS provides information on 18 indicators every 3-4 years (most recently in 2014) to monitor the goals and results of Health Population Nutrition Sector Development Programme (HPNSDP). The BDHS is a nationwide sample survey of men and women of reproductive age designed to provide information on fertility and childhood mortality levels; fertility preferences; use of family planning methods; maternal, child and newborn health, including breastfeeding practices, nutrition levels including anaemia and presence of iodine in cooking salt; knowledge and attitudes toward HIV/AIDS and other sexually transmitted infections ; and community-level data on accessibility and availability of health and family planning services. The wealth of demographic and health data that BDHS provides is essential and instrumental in monitoring and evaluating the performance of HPNSDP.

The sampling frame was the Population Census of the People's Republic of Bangladesh, clustered by Enumeration Areas (geographic areas consisting of a convenient number of dwelling units serving as counting unit for the census) and stratified by area type (rural and three urban area types). Data for the demographic and health surveys are collected from the households by face to face interview.

C10.2 What data linking has there been?

After publishing, the survey reports data are available generally for research. However, prior approval for using this data is needed, and linking of datasets with other data sets was not necessary for the project. But researchers may link BDHS data with other data sets for their own interest. However, there was no systematic attempt to develop data linkage with other datasets.

[107] http://www.niport.gov.bd/

C10.3 Barriers to data linkage

C10.3.1 Institutional factors

Lack of interest of other organizations involved in public health research for matching their data or developing linkage with BDHS data (lack of organizational interest).

C10.3.2 Operational factors

Methodological barriers: Different projects have different sample frame and methodology in collecting data based on objective of the project. BDHS uses the sampling frame comprised of enumeration areas (EAs) created for the population censuses.

C10.4 What lessons should we take away from this?

1) Collaboration with data producers at organizational level needs to be established and strengthened for the better use of data.
2) Creating awareness among the researchers about the benefits of linked data research is one of the most important factors of sharing of data in a safe and secured manner for future data linkage development.

C10.4.1 Information source

Subrata K. Bhadra is the Sr. Research Associate of National Institute of Population Research and Training (NIPORT).

C11. Data linking at the Bangladesh Bureau of Statistics

C11.1 What is the base situation?

The Bangladesh Bureau of Statistics (BBS) was formed from the merger of four other agencies in 1974. Since its inception, its role has been to provide the government of the day and the nation as a whole with statistical information to guide decision making and the development process. A key function of the BBS is to conduct the decennial Economic and Population Censuses, and make inter-Censal estimates. Other functions include methodological and geographical development work, ongoing economic and agricultural statistics, and monitoring the condition of women and children. The BBS is also tasked with preparing the National Strategy for the Development of Statistics.

The MSVSB (Monitoring the Situation of Vital Statistics of Bangladesh) project has been running in Bangladesh since 1980. The whole population is classified into clusters consisting of 100-200 households to form the primary sampling unit (PSU); through stratified sampling 2012 are taken for the MSVSB. The main objectives of the project are:

- to make population projections in the inter-censal period;
- to strengthen the existing database of vital statistics;
- to compile Demographic & Health Statistics;
- to monitor the progress of Millennium Development Goals.

The project collects data on vital events, such as births, deaths, marriages, divorces/separation, in-migration, out-migration, contraceptive use, disability & HIV/AIDS through two independent systems. Under System-1, one female local registrar is engaged in each PSU to collect data on the occurrences of the vital events in the prescribed schedules. Under System-2, staff members from district and upazila (sub-district) statistical offices collect the same data on a retrospective basis for last 3 months. Now, the responsibility has been transferred to the Deputy Directors' who performs this with the assistance of the staff members of the district office and upazila offices. Having the filled up questionnaire from two systems, data are matched at headquarters by pre-designed matching criteria and demographic rates, and ratios are calculated by the Chandra-Shekaran and Deming method.

C11.2 What data linking has there been?

Data generated from MSVSB project is available once the report is published. However, although matching is carried out to create MSVSB data itself, linking of datasets with other data sets was not necessary for the project.

C11.3 Barriers to data linkage, and how they were overcome

C11.3.1 Institutional Factors:

There is a lack of interest of other organizations involved in public health research for matching their data or developing linkage with MSVSB data; and further matching is not one of the functions of the BBS.

C11.3.2 Operational Factors:

Different statistical projects in Bangladesh have different sample frames and use different methodologies in collecting data, based on the objectives of the project. The MSVSB project has a unique methodology and sample frame to collect and collate data (methodological restriction), making it hard to link to other data sources.

Prior approval and payment is mandatory to use these data, which restricts data accessibility for the research community

C11.4 What lessons should we take away from this?

1) Data availability and accessibility needs to be improved. In a low-income country like Bangladesh, paying for data is a major hindrance in data accessibility. Data collected for the benefit of the people should be publicly available.
2) Collaboration with data producers at organizational level needs to be established and strengthened for the better use of data.
3) Creating awareness among the researchers about the benefits of linked data research is one of the most important factors of sharing of data in a safe and secured manner for future data linkage development.

C11.5 Information source

Mr AKM Ashraful Haque is one of the Deputy Directors of Bangladesh Bureau of Statistics (BBS). He is Project Director for the MSVSB project.

This work is © the Wellcome Trust and is licensed under Creative Commons Attribution 2.0 UK.

We are a global charitable foundation dedicated to achieving extraordinary improvements in human and animal health. We support the brightest minds in biomedical research and the medical humanities. Our breadth of support includes public engagement, education and the application of research to improve health. We are independent of both political and commercial interests.

Wellcome Trust
Gibbs Building215 Euston Road
London NW1 2BE, UK
T +44 (0)20 7611 8888
F +44 (0)20 7611 8545
E contact@wellcome.ac.uk
www.wellcome.ac.uk

The Wellcome Trust is a charity registered in England and Wales, no. 210183. Its sole trustee is The Wellcome Trust Limited, a company registered in England and Wales, no. 2711000 (whose registered office is at 215 Euston Road, London NW1 2BE, UK).