Cedar

Healthcare Technology Research Centre

Data Linkage Toolkit

Information about conducting projects that require linkage of health and social care datasets or access to data from established linkage repositories

Authors: Ruth Poole (Cedar)
Megan Dale (Cedar)
Daniel Thayer (SAIL)
Dr Grace Carolan-Rees (Cedar)

Date: 13 February 2015

Version: 1.3 (Full version)

GIG CYMRU NHS WALES
Bwrdd Iechyd Prifysgol Caerdydd a'r Fro
Cardiff and Vale University Health Board

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Contents

Toolkit Introduction

This toolkit has been prepared to provide information about projects that require linkage of datasets or access to established linkage repositories. It is based upon lessons learned during the CALON (Cardiac Ablation: Linking Outcomes for NICE) project in 2013-2015, introduced on page 8. Items are presented in the general order in which they are likely to be encountered, though it is often an iterative process and there may be some project-specific variations. Figure 1 summarises the main steps of a data linkage project.

Key details and recommendations are highlighted throughout in **bold** print. Lists of abbreviations and acknowledgements are included for reference.

What is data linkage?

Records containing administrative data about individuals are kept by many organisations including hospitals, GPs, social services, and even supermarket loyalty schemes. Usually, most of these routinely-collected data remain ‘in-house’, being used by the same organisation for its own purposes.

There is an increasing recognition of the value of linking records from different sources, as the combined data can provide useful information about a particular population that would not otherwise be available. The associations that are made can reveal relationships, patterns and trends that may not have been previously recognised or verified. It is possible to connect records that relate to individual people, and methods have been developed to allow this to be achieved whilst maintaining privacy. This opens up a wide range of opportunities for research and statistical analysis, with the potential to eventually improve the health and wellbeing of the population.

In this toolkit, we refer to a single collection of data as a **dataset**. Those organisations which are custodians of a dataset are known as **data providers**, as they are providing the data that are being linked together. We describe the final users of the linked dataset as **researchers**.

Animation

The ScottisH Informatics Programme (SHIP, now Farr Institute @ Scotland) produced a short animated video clip that provides a helpful introduction to data linkage in lay terms. This is available at www.scot-ship.ac.uk/public-interest. Please note that the terms they use differ from those used in this toolkit:

SHIP animation	Data Linkage Toolkit
Data custodian	Data provider
Indexing service	Trusted third party (TTP)

Whilst the procedure described to seek approval for research using SHIP may differ from those processes operated by other data providers, this animation may be helpful in understanding the basic principles of data linkage as a research methodology.

Summary of key steps

Figure 1 shows the approximate sequence of tasks undertaken in a project which is linking new data for research purposes. In reality the order is not linear but is an iterative process, with tasks being revisited or running concurrently. The timescales indicated can vary substantially, depending on a number of factors; we have provided suggested minimum timescales assuming that no major issues arise. Stages marked with an asterisk depict external dependencies (beyond the influence of the project team) that have the potential to introduce considerable delays.

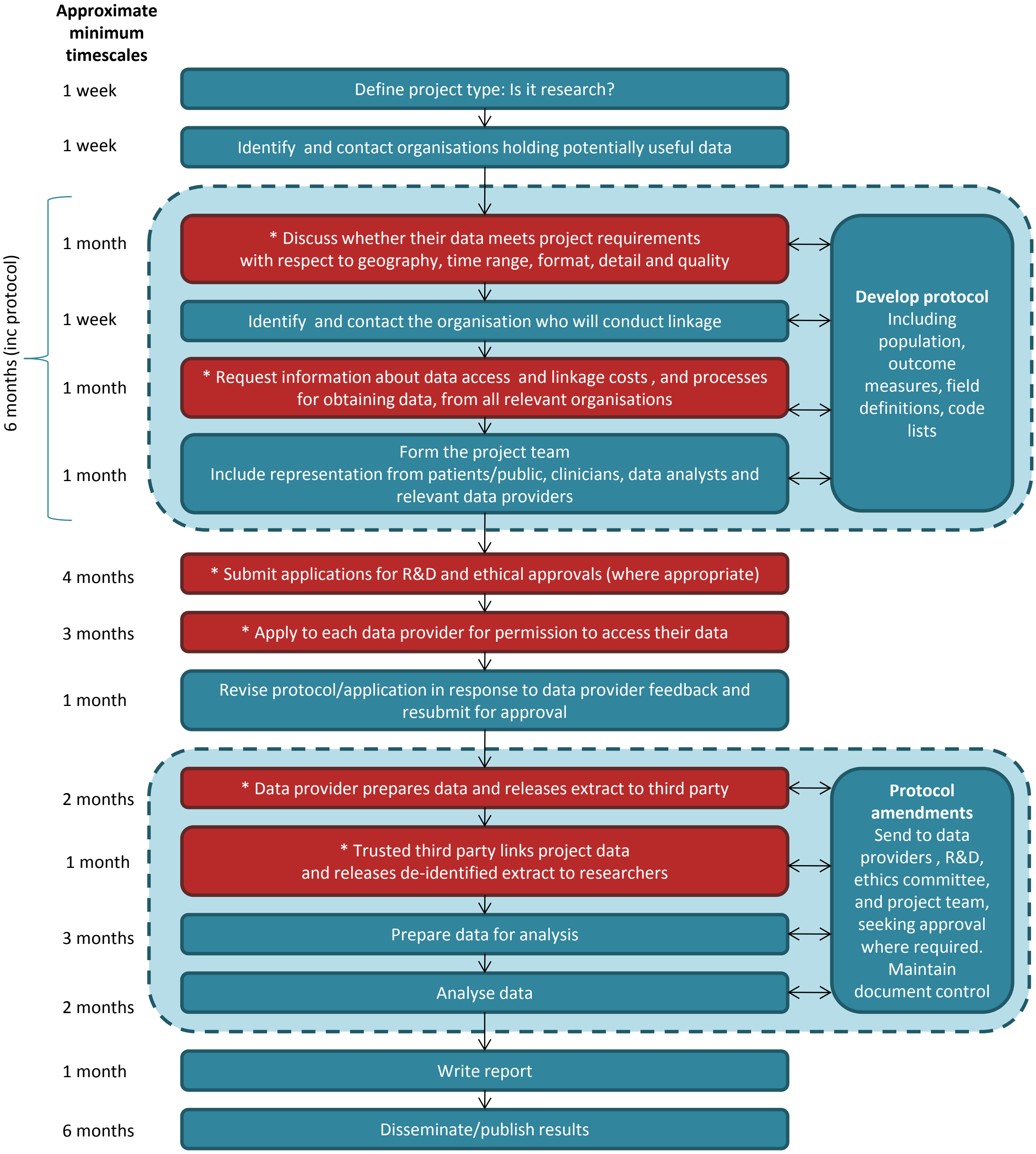

Approximate minimum timescales
1 week
Define project type: Is it research?
1 week
Identify and contact organisations holding potentially useful data
6 months (inc protocol)
1 month
* Discuss whether their data meets project requirements with respect to geography, time range, format, detail and quality
1 week
Identify and contact the organisation who will conduct linkage
1 month
* Request information about data access and linkage costs , and processes for obtaining data, from all relevant organisations
1 month
Form the project team
Include representation from patients/public, clinicians, data analysts and relevant data providers
Develop protocol
Including population, outcome measures, field definitions, code lists
4 months
* Submit applications for R&D and ethical approvals (where appropriate)
3 months
* Apply to each data provider for permission to access their data
1 month
Revise protocol/application in response to data provider feedback and resubmit for approval
2 months
* Data provider prepares data and releases extract to third party
1 month
* Trusted third party links project data and releases de-identified extract to researchers
3 months
Prepare data for analysis
2 months
Analyse data
Protocol amendments
Send to data providers , R&D, ethics committee, and project team, seeking approval where required. Maintain document control
1 month
Write report
6 months
Disseminate/publish results

Figure 1 Summary of key steps in a data linkage project. Timescales are approximate and can vary considerably. Red boxes and text marked with an asterisk indicate key external dependencies.

Introduction to the CALON project

Throughout the toolkit we include information about our experiences and lessons learned with the CALON project. This is mainly presented in shaded boxes, as seen in this general introduction below:

CALON

Cardiac Ablation: Linking Outcomes for NICE

In 2013, the National Institute for Health and Care Excellence (NICE) commissioned Cedar to design and conduct a pilot study to investigate the feasibility and usefulness of data linkage between a national register and other relevant data sources to capture health and social outcomes. The topic of cardiac ablation procedures (for treatment of arrhythmias) was chosen for Cedar's pilot study, and was titled CALON (Cardiac Ablation: Linking Outcomes for NICE).

NICE had previously reviewed safety and efficacy evidence for nine cardiac ablation procedures, four of which were deemed as requiring special arrangements for clinical governance, consent and audit or research. It was those procedures which had inadequate evidence (in terms of quantity and/or quality) that NICE was particularly interested in investigating further, by looking at existing information recorded in routine databases.

Cedar had experience of working in the area of cardiac ablation research through our Patient Reported Outcome Measures (PROMs) studies. This work had involved input from the National Institute for Cardiovascular Outcomes Research (NICOR) and use of its specialist register, which included details of cardiac ablation procedures. We had therefore already established relationships with clinical specialists and a database expert at NICOR. The existence of a suitable register allowed us to investigate the opportunities and challenges associated with establishment of new links to routine clinical datasets. Discussions with consultant cardiologists provided confirmation that the planned project had potential to provide useful information and we were assured that the recorded data satisfied the characteristics required for linkage.

The Cedar team had also developed relationships with health informatics experts at Swansea University, who are now part of Cedar's consortium. The SAIL team maintain the Secure Anonymised Information Linkage (SAIL) Databank, and have extensive experience in linking healthcare and administrative data.

In many ways the project was atypical, as our starting point was a methodology (data linkage), and the research questions were developed based on available data possibilities. Under normal circumstances a researcher will first define their research question(s) and then select an appropriate study design to address it.

A report describing the analytical methods and results of efficacy and safety outcome measures for cardiac ablation was submitted to NICE in December 2014 (Poole et al. 2014). This toolkit instead focuses on lessons learned throughout the project.

Abbreviations

Abbreviation	Full title
AALHD	Advanced Analysis of Linked Health Data
ABPI	Association of the British Pharmaceutical Industry
ADLS	Administrative Data Liaison Service
ADRN	Administrative Data Research Network
ALF	Anonymous Linking Field
AQMeN	Applied Quantitative Methods Network
CAG	Confidentiality Advisory Group
CALON	Cardiac Ablation: Linking Outcomes for NICE
CASS	Courses in Applied Social Surveys
CHI	Central Health Index
CIPHER	Centre for Improvement in Population Health through E-records Research
CPRD	Clinical Practice Research Datalink
CRM	Cardiac Rhythm Management
CSD MR UK	Cegedim Strategic Data Medical Research United Kingdom
CTV3	Clinical Terms Version 3
DPA	Data Protection Act
DSA	Data Sharing Agreement
DWP	Department for Work and Pensions
ehi^2	eHealth Industries Innovation Centre
EHR	Electronic Health Record
ESRC	Economic and Social Research Council
GP	General Practitioner
GPRD	General Practice Research Databank
HeRC	Health eResearch Centre
HES	Hospital Episode Statistics
HIRU	Health Information Research Unit
HIU	Health Informatics Unit
HMRC	Her Majesty’s Revenue & Customs
HQIP	Healthcare Quality Improvement Partnership
HRA	Health Research Authority
HSCIC	Health and Social Care Information Centre
HSCN	Health and Social Care Number
IALHD	Introductory Analysis of Linked Health Data
ICD	International Classification of Diseases
IPDLN	International Population Data Linkage Network
IRAS	Integrated Research Application System
IT	Information Technology
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NICOR	National Institute for Cardiovascular Outcomes Research
NIHR	National Institute for Health Research
NISCHR	National Institute for Social Care and Health Research
NWIS	NHS Wales Informatics Service
ONS	Office for National Statistics
OPCS	Office of Population Censuses and Surveys
PDS	Personal Demographic Service
PEDW	Patient Episode Database for Wales
PopData	Population Data BC (British Columbia)
PROMs	Patient Reported Outcome Measures
QoF	Quality and Outcomes Framework
R&D	Research and Development

Abbreviation	Full title
RCT	Randomised Controlled Trial
REC	Research Ethics Committee
SAIL	Secure Anonymised Information Linkage
SHIP	ScottisH Informatics Programme
SNOMED CT/RT	Systemized Nomenclature of Medicines – Clinical Terms/Reference Terminology
SPSS	Statistical Package for the Social Sciences
SURE	Support Unit for Research Evidence
THIN	The Health Improvement Network
TTP	Trusted Third Party
TURP	Transurethral Resection of the Prostate
UCL	University College London
UK	United Kingdom
UWA	University of Western Australia
WDS	Welsh Demographic Service

Acknowledgements

A number of organisations and individuals have been instrumental in facilitating the CALON project and the development of this toolkit. We would particularly like to thank the following for their contributions.

Organisation	Individuals
Aneurin Bevan Health Board	Alastair Roëves
Arrhythmia Alliance	Jo Jerome
Cardiff & Vale University Health Board	Maureen Fallon, Emma Lewis, Racheal James
Cardiff University	Chris Poole, Nick Francis
Cedar	Kathleen Withers, Helen Morgan
Cedar Consultant	Tony Wilkes
Clinical Practice Research Datalink (CPRD)	Jon Ford, Kendal Chidwick, Tarita Murray-Thomas, Rachael Boggon, Alan Barcroft
Department for Work and Pensions (DWP)	Nicky Tarry
Farr Institute @ London (UCL)/CALIBER	Spiros Denaxas
Health and Social Care Information Centre (HSCIC)	Garry Coleman, Diane Pryce
Healthcare Quality Improvement Partnership (HQIP)	
National Institute for Cardiovascular Outcomes Research (NICOR)	David Cunningham, Richard Schilling, Chris Gale, Polly Mitchell, Julie Sanders, Lucia Gavalova
National Institute for Health and Care Excellence (NICE)	Hannah Patrick, Phil Pugh, Sarah Garner
NHS Wales Informatics Service (NWIS)	Richard Burdon, Helen Dennis, Katy Wilson
Queen Elizabeth Hospital, Birmingham	Mauro Lencioni, Michael Griffith
ResearchOne	Samantha Crossfield
SAIL Databank/Farr Institute @ CIPHER, Swansea University	David Ford, Caroline Brooks, Daniel Thayer, Arfon Rees, Ting Wang, Luca Ruschetti, Julie Kennedy, Cynthia McNerney, Rohan Dsilva, Rod Middleton
The Health Improvement Network (THIN)	Harshvinder Bhullar

Section A – Project management

1 Introduction

Our aim in this section is to alert the reader to project management considerations that are of particular relevance to data linkage projects. The focus therefore is mainly on project planning (timescale) considerations and communications.

2 Project planning

Figure 1 in the main introduction lists common stages of data linkage projects, and could be used to inform the development of a project plan. A number of additional factors should be taken into account when estimating timescales for completion of each stage, examples of which are provided. We describe the bearing that project complexity, current context and experience might have on project progress.

2.1 Project complexity/application processes

There is likely to be a correlation between project complexity and time required, which is most evident at three stages:

- Applications to data providers
- Protocol development and data specifications
- Data preparation and analysis.

The experience of the project team may influence the amount of dedicated time needed to complete the work. Those who have previously worked together, and had prior contact with the relevant external organisations, should find that communication and application processes are more straightforward. Using familiar methodologies (with an understanding of their limitations) can facilitate preparation of study design and forestall problems. Even so, experienced researchers with established contacts have reported significant delays in accessing data from the four countries in the UK (Dattani et al., 2013).

2.1.1 Applications to data providers

Data providers generally require completion of a standard form, and some request a copy of the project protocol. Although much of the content of these forms will be similar, there is also variation in the information (and the presentation style of that information), that each organisation requests. More details about this can be found in Section H.

The approval processes of data providers vary in both length and structure, and should be taken into account when planning. **It is advisable to contact these organisations early to request details of their processes and anticipated timescales**. Whilst some are publicly available (for example from websites), we found that others are less transparent. As each project has different requirements, some data providers are reluctant to define a standard process with typical time or cost estimates.

If new data are to be linked, a substantial amount of time may be needed to negotiate data sharing agreements, as each organisation seeks assurance that the data will be protected and used in an appropriate manner. This is especially true where the data are to be linked into a permanent data linkage repository; these discussions may take years. On the other hand receiving permission to access one dataset alone, or an established data linkage repository, should be a more straightforward and comparatively less protracted process. More information about the differences between a data linkage repository and a temporary bespoke link can be found in Section B.

Timescales for approvals may also depend upon the workload of the data providers and so vary at different times of year; obtaining current estimates may be of value. Anecdotal evidence suggests that applications may be approved within a shorter timescale if repeat applications are made by organisations or individuals known to the data provider, especially where the researchers have demonstrated responsible use of data in the past.

In general, **application processes are likely to take longer if:**

- **Multiple organisations are involved**
- **A data linkage repository is being created**
- **New (external) data are being linked to an existing dataset**.

Fortunately progress is being made to streamline access to data through organisations such as the Farr Institute (see Section C for more information).

2.1.2 Protocol development and data specifications

Details of what is required at this stage can be found in Section I and Section J.

Particular challenges are faced in using observational data, which were often originally collected for purposes other than research. Depending on the ambitions of the work and the research questions being addressed, a certain level of effort will be needed to define the population(s), intervention(s), comparator(s), outcome measures and confounders. If expertise does not already exist within the project team, advisers may need to be consulted for their knowledge about the topic area, related data, or health informatics.

Sufficient time should be allocated to produce clear and comprehensive operational definitions for each data field or item. To provide assurance of scientific integrity such activities should be conducted prior to obtaining project data. It is also preferable to finalise definitions prior to submission of the protocol and/or applications for access to data, to avoid having to amend the protocol or plans at a later date. Paradoxically, it can be very difficult to accurately define the project needs without having first being able to view the type and quality of the actual data that are intended to be used. In such instances it may be possible to ask for a sample or to visit the data provider to gain further insight.

2.1.3 Data preparation and analysis

Preparation of data in a form suitable for analysis is likely to be an iterative process, where researchers and informaticians work together in cleaning and organising the project dataset (see Section K for more information). Whilst operational definitions and code lists will have been produced when the protocol was written, further decisions may be needed later on about the details of data matching, application of exclusion criteria, and formatting prior to analysis. Unanticipated queries might arise that lead to further refinement of definitions. Even once the dataset is ready for use, the data will need to undergo statistical analysis, perhaps including the preparation of tables and graphs. Furthermore, if the data are stored and analysed within a secure environment, there may be information governance processes that are required before the outputs are released. All of these activities contribute to the duration of this stage of the project.

2.2 Current data protection climate

In planning a project that makes use of individual patient data, consideration should be given to current legislation and local attitudes to data protection.

2.2.1 United Kingdom

In 2013, NHS England commissioned a programme of work that involved linkage of health and social care data at a national level, and ultimately aimed to use the processed information to improve patient care. The initiative was named care.data. Collection and linkage of these data was to be managed by the Health and Social Care Information Centre (HSCIC). Early in 2014 the media began highlighting concerns about the extraction of data from primary care electronic records for these purposes. There has since been widespread criticism and detailed scrutiny of HSCIC's methods.

HSCIC responded by temporarily suspending the release of data and all their linkage activities for several months whilst they made efforts to address these concerns. This suspension has since been lifted and HSCIC has been working to manage the backlog of data requests. An information dashboard is produced by HSCIC, showing how long it takes for them to process requests. In February 2015 HSCIC announced that the backlog had been cleared, and that applications would take between 14 and 60 days to process, depending on complexity. Any **researchers hoping to use hospital data from England and/or obtain linked primary care data from the Clinical Practice Research Datalink (CPRD) may wish to confirm the current status**.

CALON

During the course of the CALON project, events occurring at a national level as a result of care.data impacted significantly on the project schedule and study design. The suspension of HSCIC's linkage activities meant that none of the primary or secondary care records from England were available for CALON within the project timescale. As HSCIC were also responsible for provision of secondary care data from English hospitals, there would have been no advantage in seeking alternative linkage services with another organisation.

In this instance, no-one had anticipated the magnitude of the impact of the care.data programme launch. This could not have been planned for at the outset. The one mitigating factor in CALON was that only part of the project relied upon data from England; data from Wales were linked and stored within the SAIL Databank, and not managed by HSCIC.

Lessons learned are:

- **Consider whether alternative data sources are available if any problems occur with the organisation(s) you plan to use**
- **Be aware of recent developments and attitudes towards data protection issues.**

2.2.2 European legislation

In 2012 the European Commission proposed that the EU's Data Protection Directive (95/46/EC) is replaced by a General Data Protection Regulation (European Commission 2012). Under the current directive, the European data protection rules could be adapted to suit each member state. The Regulation however would be less flexible, being directly applicable and consistent throughout the EU (Ploem et al. 2013). At the time of writing this report, the Regulation has not yet come into force, but data protection reform has been a certainty since the European Parliament voted to support it in March 2014 (European Commission 2014).

The draft proposal includes the following paragraphs:

41) Personal data which are, by their nature, particularly sensitive and vulnerable in relation to fundamental rights or privacy, deserve special protection. Such data should not be processed, unless the data subject gives his explicit consent. However derogations from this prohibition should be explicitly provided for in respect of specific needs, in particular where the processing is carried out in the course of legitimate activities by certain associations or foundations the purpose of which is to permit the exercise of fundamental freedoms.

42) Derogating from the prohibition on processing sensitive categories of data should also be allowed if done by a law, and subject to suitable safeguards, so as to protect personal data and other fundamental rights, where grounds of public interest so justify and in particular for health purposes, including public health and social protection and the management of health-care services, especially in order to ensure the quality and cost-effectiveness of the procedures used for settling claims for benefits and services in the health insurance system, or for historical, statistical and scientific research purposes (European Commission 2012).

Effectively, the proposal allows the processing of data for historical, statistical or scientific research without the need for consent or another legal basis, provided it satisfies the requirements of Article 83.1, which states:

> "Within the limits of this Regulation, personal data may be processed for historical, statistical or scientific research only if:
> a) these purposes cannot be otherwise fulfilled by processing data which does not permit or not any longer permit the identification of the data subject;
> b) data enabling the attribution of information to an identified or identifiable data subject is kept separately from the other information as long as these purposes can be fulfilled in this manner" (European Commission 2012).

Amendments to the draft regulation were later proposed, some of which raised concerns amongst medical research organisations, as they might have prevented research based upon individual medical records unless patients had explicitly consented to the use of their data for that specific purpose. The draft regulation and amendments continue to be debated by various committees, but these negotiations are expected to be completed by the end of 2015 (European Commission, 2015).

Implementation is not anticipated before 2017. **Those managing projects that make use of observational data** (and are likely to extend into the medium to long term) **are advised to monitor this situation and consider the implications for their work**.

3 Communications

Irrespective of the project design, it is unlikely that one individual (or even organisation) will possess all the resources, skills and knowledge needed. **Effective communication is therefore crucial to the success of data linkage projects**. Section D provides a suggested list of stakeholders and their roles, including the steering group.

3.1 Project team, steering group and newsletters

The core project team may be employed by more than one organisation. Due to the complexity of CALON, we found it very helpful to hold regular internal progress meetings and record key discussion points.

The CALON steering group was set up to guide the project, provide specialist information and insights, and ensure that all relevant issues are covered from the perspective of each member. Regular contact has been made with members through group meetings and individual consultation with topic experts where appropriate. Their personal and collective input was highly valued at all stages of the work.

In addition to this, occasional newsletters were circulated. They informed a wider group of project stakeholders who didn’t need to be involved in the details, but who may have been interested in hearing about the ongoing work.

3.2 External contacts

Efforts taken to identify and engage the most appropriate and helpful external individuals could be key to the success of the entire project. With the CALON project, we had first established which organisations had the potential to provide the data or services we required (see Section E).

Our experiences of achieving our communication goals (whether obtaining information, arranging meetings, or simply establishing a relationship) varied considerably, particularly with respect to data providers. We tried a number of different approaches – emailing individuals directly, emailing generic helpdesks, telephoning main contact numbers, visiting offices, and even approaching individuals at conferences/events. We also followed up recommendations from fellow researchers.

We were unable to conclude that any single approach is ideal, as people and organisations vary in their preferred communication methods and their responsiveness. **We would advise persistence, patience and employment of multiple communication strategies**. In most cases, we were able to eventually track down a helpful individual who, if not able to address all our queries, would endeavour to point us in the right direction.

Within some organisations people may work at different geographical locations or organisational departments, with limited internal communications. **It is best not to assume that details of a discussion held with one individual will be known to others**.

3.3 Ongoing monitoring

As with all complex projects with multiple contributors and several contemporaneous ongoing tasks, **strategies must be employed to deal with competing demands without detracting attention from high priority activities**. Regular progress checks and review of communications may be helpful.

Section B – Topic selection

1 Introduction

Certain research topics are more conducive to the use of routinely collected data and data linkage methodologies than others. In some cases, observational data may supplement other data collected. For example, a Randomised Controlled Trial (RCT) may collect new data on trial participants, with longer-term outcomes collected through linkage to existing datasets (such as hospital records).

In deciding whether data linkage is appropriate for a particular research aim, it is worth reviewing some characteristics that facilitate the use of these methods, and the types of project that they are commonly applied to. We conclude this section with a summary of some of the benefits of using linked data for research purposes.

2 Characteristics

In discussion with health informatics experts at the SAIL Databank, we identified a number of characteristics that might make a topic suitable for investigation through data linkage:

- **Sufficient numbers of patients**
    - For protection of patient identities, especially where there are sub-group analyses (most data providers do not allow publication of results in cells that refer to a small number of individuals, such as less than five).
    - Make sure that adequate numbers are likely to have been recorded in the UK. Some data providers will conduct a 'feasibility' exercise to check this in advance before resources are committed to a full-scale study.
- **Data that have been recorded for at least a few years**
    - There may be a lag in time between the intervention or outcome and the date it is recorded. Furthermore there is often a delay between data entry and the time at which the data are released from providers, to allow some checking/cleaning to take place.
    - The longer the period of data capture, the more individuals are likely to be included.
    - Be aware of historical changes to the data, such as introduction of new classification systems, alterations to the minimum dataset, and changes to the intervention or health care services.
- **Data available from appropriate UK datasets within a reasonable length of time**
    - Some types of data are known to be difficult or slow to obtain, such as mortality data from the Office for National Statistics (ONS).
    - See Section A and Section H for more information about factors affecting time required to access data.
- **Intervention and outcome data that are likely to be well recorded**

- In routine clinical data, some types of data are better recorded than others. For example, evidence of prescriptions can usually be found in GP records (though may lack some details). More information about data considerations can be found in Section J.
- Patient symptoms are not often recorded unless the dataset was designed for this purpose, for example in a Patient Reported Outcome Measures (PROMs) or patient satisfaction questionnaire study.

- **Identifiers within each dataset to allow for matching of records**
  - Matching individual patient records from separate sources relies on accurately identifying the same person in both datasets. The ideal identifiers are unique to each individual (for example NHS number or National Insurance number).
  - Where unique identifiers are not available, matching may still be possible using a combination of other variables such as date of birth, postcode and sex. A description of the linkage process and how records are matched can be found in Section F.
  - Appropriate information governance measures must be in place (see Section G).
- **Data that have the potential to be used to address a relevant research question**
  - A typical project will already have defined the research question(s) before designing the study methodology. If seeking to answer those questions using routine data, it is important to make sure that the correct data have been recorded, and that it is in a suitable form for analysis.
  - The CALON project was atypical as we defined the research questions after deciding on a methodology. The aim was to provide additional efficacy and safety evidence for one or more cardiac ablation procedures identified by NICE as requiring further research. We attempted also to look for social outcomes that could be evaluated alongside clinical data, such as time off work.

3 Applications of data linkage methodologies

In order to understand why data linkage methodologies are more appropriate for certain types of research than others, it is helpful to first consider the differences between ongoing data linkage repositories and ad hoc ‘snapshots’ of data that have been compiled through bespoke links.

3.1 Repository versus snapshot

Data can be linked on an ad hoc basis, where links are set up for a specified purpose and a distinct period. Only those datasets and variables required to address the predefined research objectives are requested. Alternatively, an ongoing data linkage repository may be developed, incorporating multiple datasets to support an indefinite number of projects.

Data linkage repositories are ultimately more likely to provide the maximum benefit to healthcare research as a whole. Multiple interrogations of these reusable data are possible, leading to improvements in data quality as researchers all contribute to checking/cleaning of data and creation of algorithms to define variables. These systems are utilised by specialist e-health research centres, such as the Centre for the Improvement of Population Health through E-records Research (CIPHER) who use the SAIL Databank (see Section E for information about data resources).

There are a number of challenges in building such repositories. They require dedicated funding to support a central ongoing unit, with massive computing capacity and staff with specialist processing and analytical knowledge and skills. In contrast, smaller scale ad hoc data linkages are created for a defined period and can draw upon research grants for specific projects. The most efficient means of accessing data (in terms of resource use) for a single project is to make use of those data that are already available within a comprehensive, linked repository.

CALON

One of the aims of the CALON pilot study was to link registry data to other datasets, and not simply to make use of an established data linkage repository. Due to the limited time available to Cedar in conducting the pilot study, an ad hoc linkage (to an existing repository) was the only feasible option on this occasion. The alternative, creation of an ongoing link between the NICOR register and the SAIL Databank, would have required lengthy data sharing negotiations between the data providers.

In general, accessing data already held within a repository would be preferred, as the proportion of benefits to effort is likely to be much greater. The incorporation of new datasets into these repositories should be encouraged, although external researchers may find they have little influence in such processes.

Repositories are particularly advantageous when looking beyond the boundaries of an individual project to the potential collective opportunities presented by a data linkage repository to multiple researchers, and the more efficient utilisation of resources. The Farr Institute (see Section C) is facilitating collaborative working across the UK, establishing a co-ordinated approach in providing access to such data resources.

3.2 Typical research questions

On their website, the SAIL team list some examples of research questions that could be answered using the SAIL Databank:

- Is disease X increasing or decreasing?
- Might early childhood medication later affect how well they do in school?
- How does poverty affect the need and demand for health services?
- What are the long term outcomes of the Welsh government's anti-smoking policy?
- Are there enough patients suitable for a new clinical trial in a specific area in Wales?
- If care is redesigned in a particular way, what will be the likely impact on GP services and hospital services, and on different populations (for example different age groups)?
- How many patients would benefit from a new treatment (supported by NICE) and how much would this cost?

3.3 Applications and examples

A number of applications for this type of observational research can be found in the published literature. Some countries have a longer history of working with large datasets than others due to the configuration of their healthcare systems, access to suitable records, and investment in data linkage skills and infrastructure. Areas with such expertise outside of the UK include Canada, Australia and the Nordic countries.

Some examples of applications are described below, under the following headings:

- Interventions and outcomes
    - Point-of-care trials
- Social outcomes and determinants of health
- Health service utilisation
- Disease aetiology
    - Genetic and phenotypic linkages
- Disease surveillance
- Methodological development.

3.3.1 Interventions and outcomes

In contrast to medications, new medical devices and procedures are less tightly regulated and may enter into use in the NHS with relatively scant evidence for their efficacy and safety. Details of medical devices are rarely recorded in routine datasets, and as such may not be specifically identifiable.

Conducting observational research into new interventions may prove challenging; a consequence of their novelty being that routine datasets may not contain large quantities of data. However, clinical trials of these new interventions are likely to produce even fewer data and require more time, whereas routine data are available soon after treatment is administered. Interventions that are better established can provide a wealth of data for evaluation.

A comparative study by Reilly and co-workers (2012) of treatments for precancerous changes to the cervix was conducted to investigate birth outcomes (preterm birth and low birth weight) in subsequent pregnancies. The exposure groups were women who underwent colposcopy only, colposcopy with subsequent treatment, and a control group of women who had negative smear tests only. The authors concluded that women with abnormal cervical smears who were referred for colposcopy had an increased risk of preterm birth irrespective of whether they underwent treatment (excisional, ablative or other). This was the first record linkage study on this subject in the UK, and was made possible by linkage of a cervical screening database to a database of child health which recorded birth weight, gestational age and maternal factors.

Point-of-care trials

Technology which allows the recruitment of patients into trials within the context of a GP consultation has been developed. Whilst has been demonstrated to be a feasible proposition (Brooks et al. 2009), recruitment of clinicians willing to incorporate this additional work into their usual practices has proven challenging (van Staa et al. 2014).

3.3.2 Social outcomes and determinants of health

An advantage of linking data from diverse sources is that cross-disciplinary research can reveal patterns that might not otherwise have been recognised. Alongside health data, is it possible to make use of educational, employment, residential, judicial and familial data for research purposes, and there are likely to be many more avenues that have not yet been explored.

Access to both Social Services records and admissions to the burns unit of a hospital enabled the assessment of a cohort of children against a matched control group (who had not experienced a burn). The majority of burns were deemed to be accidental, but it was found that those children who had been burned were statistically more likely to have been subsequently referred to Social Services due to neglect or abuse (James-Ellison et al. 2009).

3.3.3 Health services utilisation

Linked records from a variety of sources allow an evaluation of usage of healthcare services and associated costs, whilst controlling for comorbidities and potential confounders at the level of an individual patient.

Morgan et al. (2014) measured service utilisation of pregnant women, including primary and secondary care access and prescriptions. Using these data together with econometric analyses, the authors concluded that increased health service usage and healthcare costs were associated with increasing maternal body mass index, having adjusted for other appropriate parameters.

3.3.4 Equity

The availability of large-scale population-level data enables evaluation of the equity of access to healthcare services. Similarly, the influence of socio-demographic factors on the development and prognosis of disease can be assessed.

Linkage of cancer, hospital and death records has been used to study the likelihood of women receiving breast reconstructive surgery after surgery for breast cancer. Hall and Holman (2003) reported that various factors influenced the rate of reconstructive surgery – the likelihood decreasing with age, for women from low socio-economic groups, or for those from rural areas.

By linking hospital inpatient admission and mortality data with measurements of air pollution, Roberts et al. (2012) demonstrated that incidence of serious asthma was more strongly associated with social deprivation than exposure to air pollutants.

3.3.5 Disease aetiology

Routinely-recorded data permits the tracing of associations between exposures and outcomes, sometimes even when a significant period of time has elapsed between them. The respective influence of a number of concomitant risk factors can be assessed.

Linking records of people who had undertaken long-haul flights with hospital admissions for deep vein thrombosis or pulmonary embolism, Kelman et al. (2003) were able to evaluate the magnitude of risk of venous thromboembolism after air travel. The hazard period had previously been estimated to fall within two to four weeks after flying, but the linked data revealed that the risk was in fact highest within the first two weeks.

Genetic and phenotypic linkages

Now that the Human Genome Project has allowed the sequencing of DNA to be undertaken efficiently, there is scope to link the observed trends to other types of healthcare data. Linkage of historical genealogical data with identification of genetic variants enables the traits underlying phenotypes and common diseases to be studied.

Initiatives such as the UK Biobank also provide useful data that can contribute to this type of research.

3.3.6 Disease surveillance

Routine data sources may contain historical data that can be traced back over many years. They can therefore be a valuable resource for discovering more about the natural history of diseases and factors that might influence prognoses.

Large-scale follow-up after hospital admission for Crohn's Disease and Ulcerative Colitis was achieved through linking hospital and mortality records. The study enabled calculation of prevalence of severe illness in different geographical areas and against associated potential risk factors (such as social deprivation, urban/rural environments), as well as by demographic profile (age and sex). Long-term follow-up enabled assessment of mortality rates up to five years following hospitalisation (Button et al. 2010).

3.3.7 Methodological development

Use of linked observational data for health and social care research is in its infancy in many parts of the globe, though more advanced in the UK than the majority of other locations. Few organisations have the correct combination of skills, experience and infrastructure systems to make the best use of very large datasets. This means that the field of health informatics is ripe for the development of innovative techniques and applications. Methodological developments, such as the refinement of coding algorithms, are a common goal or by-product of research of this type.

4 Benefits

Whilst there are challenges in linking and using routine administrative and healthcare data, there are also some clear benefits.

4.1 Cost-effectiveness

Linkage of existing routine data is less costly than implementation of prospective longitudinal studies or randomised controlled trials, and can produce more timely results.

4.2 Patient burden

As there is often little or no contact with individual study subjects in retrospective studies of existing data, and no intervention over and above the normal care pathway, the burden placed on patients is reduced.

4.3 Ethical

In circumstances where one treatment has clear benefits over any alternatives, it would be unethical to conduct a randomised controlled trial. However it may still be of value to evaluate the long-term outcomes of that treatment.

4.4 Scientific

As illustrated by examples provided above, one of the main benefits of this type of research is the enhancement of knowledge. Resulting policy reforms and changes to provision of services can ultimately impact upon the health and well-being of the population as a whole.

4.5 Sub-group analyses and vulnerable populations

Analysis of routine data is particularly beneficial when investigating populations who may otherwise be difficult to access, such as ethnic and socioeconomic groups who are typically under-represented in other types of research.

4.6 Protection of privacy

The design of secure data linkage repositories allows research to be conducted at an individual record level without the identity of the person being revealed. The SAIL Databank is one example of a resource that has been developed for this purpose. Jones et al. (2014) describe the design, principles, operating model and features of the SAIL Gateway, explaining how data can be remotely accessed by researchers.

4.7 Collaborative working

As described in Section A, development of data linkage projects and systems often require complex interactions between multiple organisations and effective communications. A potential consequence of this is an improvement in relationships between researchers, clinicians, analysts, patient group representatives and others.

4.8 Economic

Strengthening data linkage skills, experience, collaborations and infrastructure improves the UK’s capacity for research. This can attract further investment from sources such as large pharmaceutical companies.

Section C – Resources

1 Introduction

Research using linked administrative and healthcare data in general is growing in popularity and accessibility, though relatively few have the full skill-set and/or computing facilities required to conduct this type of work. Investment in infrastructure and training is attempting to address this need in the UK, as the field of "big data" research is rapidly escalating. Similarly, resources available to researchers are continually developing, and as such the advice contained in this section may only be helpful in the short-term as a starting point for further investigations.

In 2013 we were at the early stages of planning the CALON project. Our experience at this time was that it was not immediately clear where general guidance could be accessed. There was a lack of co-ordination of guidance within and between UK organisations. It is encouraging to see now that there has been an expansion in the volume and quality of advice available, though there is still room for improvement.

In this section, we list some of the resources that we encountered throughout the course of our project, including general data linkage organisations, conferences, mailing lists, a key report, and training opportunities. For details of some specific datasets and data providers in the UK, see Section E. Assistance that may be available from data providers in defining project parameters is described in Section J.

2 Data linkage and e-health organisations

2.1 Farr Institute of Health Informatics Research

In March 2013, four Centres of Excellence in e-health informatics research were established across the UK. The Farr Institute of Health Informatics Research supports the collective work of these centres, led by the following organisations:

- University College London (Farr Institute @ London)
- University of Manchester (Farr Institute @ HeRC N8)
- Swansea University (Farr Institute @ CIPHER)
- University of Dundee (Farr Institute @ Scotland)

The Institute aims to establish a coordinated approach to data safe havens, create digital laboratories for large scale research, widen access to well-described datasets and facilitate communication to address key issues in health informatics research. These issues include governance, computer science infrastructure, public engagement, training and education. It is expected that this work will support innovation in the public sector and industry leading to advances in preventative medicine, improvements in healthcare delivery, and better development of commercial drugs and diagnostics.

2.1.1 The College of Medicine at Swansea University

Our work on the CALON project has involved a close collaboration with colleagues from the College of Medicine at Swansea University, who have expertise in the management and use of data within the SAIL Databank. These individuals also contribute to the work of the CIPHER, part of the Farr Institute of Health Informatics Research. In 2014 this team in Swansea formally became part of Cedar's consortium, contributing to our work as an external assessment centre for NICE. As such, processes are being developed that will allow streamlining of access to data and linkage services for Cedar researchers in future.

As part of the Farr Institute, the team at Swansea are working towards becoming a 'one-stop' centre for accessing linked health data. These data will not be restricted to Wales, as linkages to CPRD, HSCIC and other data providers in England and the rest of the UK are in the pipeline.

For those new to analysis and use of large datasets, **we consider engagement with those with such experience, knowledge and skills, as invaluable in negotiating the specific complexities of this type of research**.

2.2 International Population Data Linkage Network

Formerly the International Health Data Linkage Network (IHDLN), the International Population Data Linkage Network (IPDLN) aims to facilitate communication between centres specialising in data linkage services, and users of linked data. Having initially focused on health, the network was recently renamed to reflect the broader nature of research being conducted. A conference is held every two years (see below).

The aims of the Network are to:

- Establish and maintain an effective and useful network of data linkage centres
- Foster collaboration and exchange programmes between data linkage centres
- Produce a compendium of measurements based on linkage of data across Australia, Canada and the UK
- Record the outputs from data linkage activities and programmes across the globe.

Organisations and individuals who use linked data may wish to apply for membership of the network to keep up-to-date with new developments and to share knowledge.

2.3 Administrative data services

2.3.1 Administrative Data Liaison Service

The Administrative Data Liaison Service (ADLS) was set up to support administrative data based research in the UK. The services available to researchers include an advisory service, safe researcher training, coding archives and a TTP data linkage facility.

Although its emphasis leans towards the social sciences rather than health care, much of the general advice provided on the ADLS website is equally applicable to both. A limited amount of information about "health and disability" datasets can currently be found under the theme of the same name; but is largely confined to resources from NHS Scotland and some details of Hospital Episodes Statistics (HES, hospital data from England provided by HSCIC). Some of the responsibilities of the ADLS are now being moved over to the Administrative Data Research Network.

2.3.2 Administrative Data Research Network

A relatively recent development is the creation of the Administrative Data Research Network (ADRN), funded by the Economic and Social Research Council (ESRC). The network enables research based on linked governmental data. The main Administrative Data Service is hosted by the University of Essex, with four Administrative Data Research Centres being located across the UK at the University of Southampton, Queens University Belfast, the University of Edinburgh, and Swansea University. The ADRN does not store administrative data, but can assist researchers in negotiating access to data from governmental departments on a case-by-case basis.

3 Conferences and events

3.1 Farr Institute International Conference

SHIP previously ran a biennial international conference entitled "Exploiting Existing Data for Health Research". The three-day programme presented keynote speeches, panel discussions and conference sessions offering seven parallel themes. Topics included information governance, methodological and data processing challenges, and examples of e-health research across a variety of clinical and social care subjects. The Cedar team found this an excellent forum for learning more about using linked health data and networking with experts in this field.

The conference is now run by the Farr Institute, and in 2015 is being held in St Andrews, Scotland on 26-28th August. More information about the conference can be found here.

3.2 International Population Data Linkage Conference

This conference is run by the IPDLN (see above), and was previously known as the International Health Data Linkage conference.

The next conference will be hosted by Swansea University and held on 20-22nd July 2016 in Cardiff.

3.3 Other events

Numerous meetings are being held at local, national and international levels in response to the surge in interest in "Big data" and the "Internet of Things". These opportunities vary in their usefulness, and **it is worthwhile prioritising those that align most closely with your particular research interests**.

4 Mailing lists

In such a rapidly developing research environment, it can be challenging to keep up-to-date with the latest news. At Cedar we have found it helpful to sign up to a number of email distribution lists, though the relevance of contents was variable. Organisations that send out regular mailings include:

- HSCIC – 'Data Insight' bulletin
- Office for National Statistics (ONS)
- Department of Health – Digital Health
- Royal College of Physicians, Health Informatics Unit (HIU)
- Healthcare Quality Improvement Partnership (HQIP)
- Applied Quantitative Methods Network (AQMeN)
- eHealth Industries Innovation (ehi^2) centre
- Administrative Data Research Network (ADRN).

For the CALON project, we also subscribed to newsletters from professional societies and patient organisations that were related to the clinical area of interest.

5 Big data road map

In November 2013, the Association of the British Pharmaceutical Industry (ABPI) held an event entitled "360^o of Health Data: Harnessing Big Data for Better Health". The main purpose of the day was to introduce their "Big data road map" – an overview of the status of UK health informatics capabilities and infrastructure, which also laid out a strategy for further development. Whilst presenting information from the perspective of the pharmaceutical industry, this 35-page report provides a helpful introduction to the opportunities and challenges presented by Big Data.

6 Training

Table 1 summarises some training courses, mainly focusing on data linkage and/or analysis of linked data, and information governance. A key to training providers follows. Please note that this is not an exhaustive list; appropriate resources may also be available elsewhere. Also note that the term 'administrative data' often covers a broad spectrum of data types which might include healthcare data.

Table 1 Training opportunities.

Title	**Provider/venue**	**Format**	**Content**	**Website**
Administrative Data	PopData	Online (1 hour)	Basic introduction to working with administrative data, opportunities and challenges, key resources.	www.popdata.bc.ca/etu/onlinecourses/ADMN101
Advanced Analysis of Linked Health Data	Swansea/UWA	Five-day course	Provides health and social researchers with the opportunity to build on their pre-existing theoretical knowledge and skills in the analysis of linked data by exploring a number of advanced topics.	www.swansea.ac.uk/medicine/courses/msc-health-informatics/analysisoflinkedhealthdata/advancedanalysisoflinkedhealthdata/
Analysis of Linked Datasets	ADRCE/Southampton	Two-day introductory course	Data linkage procedures and the analysis of linked datasets subject to linkage errors.	http://www.ncrm.ac.uk/training/show.php?article=5084
Combining Data from Multiple Administrative and Survey Sources for Statistical Purposes	ADRCE/Southampton	Three-day introductory course	Focus on statistical techniques and understanding the origin and nature of potential errors found in integrated datasets.	http://store.southampton.ac.uk/browse/extra_info.asp?compid=1&modid=5&deptid=39&catid=113&prodid=576
Data Linkage: From Theory to Practice	NCRM/ADRCE	Three-day course	More intensive than the introductory course, this introduces concepts and methods of record linkage and evaluation of techniques.	http://store.southampton.ac.uk/browse/extra_info.asp?compid=1&modid=5&deptid=39&catid=113&prodid=479
Harnessing Electronic Health Records for Research	Farr Institute	Programme of 13 inter-related short courses (1 or 2 days)	Topics include "Answering Clinical Research Questions with Health Records" and "National Registries: From Audit to Research".	http://www.ucl.ac.uk/farr-short-courses

Title	Provider/venue	Format	Content	Website
Information Governance	SHIP	Online (approx.. 15 hours)	Legal concepts involved in secondary use of health data, information governance, statistical disclosure control, data security and data protection.	http://www.law.ed.ac.uk/teaching/online_distance_learning/cpd_courses/ship_information_governance
Introduction to Data Linkage	ADRCE/Farr Institute/Southampton	One-day introductory course	Uses of data linkage, data preparation, methods for and issues for the analysis of linked data. Focus on health data.	http://store.southampton.ac.uk/browse/extra_info.asp?compid=1&modid=5&deptid=39&catid=113&prodid=480
Introduction to Linking Data	CMIST	One-day introductory course	Basic concepts of data linkage, data linkage applications, data sources, preparing datasets for data linkage.	http://www.cmist.manchester.ac.uk/study/courses/short/introductory/intro-to-linking-data/
Introductory Analysis of Linked Health Data Swansea/UWA	Swansea/UWA	Five-day introductory course	Theory and practice of analysis of large sets of linked health and social data. Epidemiological principles and computing concepts.	www.swansea.ac.uk/medicine/courses/msc-health-informatics/analysisoflinkedhealthdata/introductoryanalysisoflinkedhealthdata/
MSc Health Data Science	Swansea	One-year full-time (three-years part-time) taught master's	Develops skills and knowledge in processing health data to extract information about individuals and populations.	http://www.swansea.ac.uk/postgraduate/taught/medicine/msc-health-data-science/
Safe Researcher Training	ADLS	One-day course, various locations	Focus on safe and responsible usage of UK administrative data, data security, good practice and statistical disclosure control.	http://www.adls.ac.uk/safe-researcher-training/

Key:

ADRCE	Administrative Data Research Centre England
ADLS	Administrative Data Liaison Service
CMIST	Cathie Marsh Institute for Social Research, University of Manchester
PopData	Population Data BC, University of British Columbia, Vancouver, Canada (online course accessible from UK)
SHIP	ScottisH Informatics Programme
Southampton	University of Southampton
Swansea/UWA	Swansea University (course location), training provided by University of Western Australia
Swansea	College of Medicine, Swansea University
UCL	Institute of Child Health, University College London

Section D – Project stakeholders

1 Introduction

It is highly unlikely that any one individual will possess all the knowledge, experience and abilities needed to successfully complete this type of project alone. Bringing together appropriate individuals early in the project is a key consideration.

Analytical abilities are needed for data preparation, processing and statistical analysis. Additionally, some understanding of each dataset and its classification and coding systems informs the way that the data might be used. Those with insight into the clinical aspects of disease and treatments should be involved in the study design and interpretation of results. This would apply to both medical professionals and patient representatives, who might provide different perspectives on health states and the value of interventions or outcomes. Coordination of all of these contributions requires project management skills, and guidance from those commissioning the work helps to define the scope and relevance.

Throughout this section we describe the involvement of some of the stakeholders who have played a part in CALON, as an example of the types of roles that might contribute to similar projects. After listing the steering group participants, we refer to others who provided expertise at different stages of our work. Named individuals are acknowledged in the main toolkit Introduction.

2 Steering group

The steering group's purpose in CALON was "To guide the project, provide specialist information and insights, and ensure that all relevant issues are covered from the perspective of each member". Information about our communications with the group can be found in Section A. The group was comprised of a number of people with the following roles:

- Project oversight and management (Cedar)
- Commissioner's representative (NICE)
- Data analysts (SAIL)
- Experienced researchers (NICOR; SAIL; Cardiff University; Cedar)
- Patient group representative (Arrhythmia Alliance)
- Clinical representatives (various).

2.1 Oversight and project management

The Cedar team have expertise in health technology evaluation and project management, and all our researchers are registered PRINCE2® Practitioners. The Cedar Director has overall responsibility, but day to day project management is led by a Cedar researcher. In addition to the project lead, a second team member was available for assistance and quality checking.

The complexity of the data linkage process increased the importance of good project management and communication. It was **helpful to have:**

- **Key team members leading the project throughout the entire process**, to maintain continuity
- Documentation of project plans, history, risks, progress and major decisions.

Delegation of some activities and responsibilities to others outside of Cedar helped to balance workloads and maximise utilisation of specialist skills.

2.2 Commissioner's representative

The National Institute for Health and Care Excellence (NICE), who commissioned CALON, also allocated its own project lead to provide guidance and monitor progress. Their input to the steering group was crucial, providing useful suggestions and ensuring that planned activities were in harmony with the desired outputs from NICE's perspective.

2.3 Data analysts

The SAIL team at Swansea University are part of Cedar's consortium, and employ dedicated staff to support Cedar's work. Their analysts work with data from the SAIL Databank, and conduct linkage to other external datasets. They also have experience of setting up and maintaining complex data collection systems through specialist registries, patient questionnaires and bespoke clinical information extracts.

This **health informatics expertise was crucial to the success of CALON**. The SAIL team supported all aspects of the project, but their skills came to the fore in incorporating new data from the (external) specialist registry, linking the records and helping to prepare them for statistical analysis.

2.4 Experienced researchers

Researchers from organisations with experience in using particular datasets were invited to contribute to the project. The purpose, structure and content of each dataset varied considerably, and it was helpful to have input from those who have used them previously for research. **Naming such individuals may also strengthen applications for data, especially where publications are available that demonstrate responsible use of data in the past**.

2.5 Patient group representative

To represent the views of patients with arrhythmias and their carers, a spokesperson from the Arrhythmia Alliance was invited to sit on the steering group. Their contributions throughout the project were invaluable, particularly in defining research questions and understanding the implications of results from the perspective of patients with arrhythmias.

2.6 Clinical representatives

CALON steering group members included a cardiac ablation specialist nurse, two Consultant Cardiologists, and the Primary Care Clinical Director of a local health board (who also works as a General Practitioner). As well as understanding the typical needs of patients, care pathways, available treatments and expected outcomes, these clinicians collectively had prior experience of using NICOR's specialist register, PROMs and primary care data. They were therefore able to assist in defining procedures and outcomes using various coding classification systems (see Section J for more information about coding), and helped to interpret the project results within the relevant clinical contexts.

3 Other contributors

Whilst not all represented on the steering group, a number of other individuals and organisations also made significant contributions to CALON.

3.1 Coding experts

As the task of defining procedures and outcomes relies heavily on the coding classification systems used, we asked senior clinical coding experts for their assistance. The Classification Standards Manager from the NHS Wales Informatics Service (NWIS) kindly advised on factors to consider in designing the study, and helped to produce lists of codes to search for in the project dataset.

3.2 Statistical advice

Cedar's statistical consultant helped to plan appropriate analyses for the efficacy and safety data, and to interpret the results. Analysis of large, complex datasets based upon routine administrative data can present particular challenges that may be less problematic in smaller, more tightly controlled research studies; one example being potentially high volumes of missing data. Other experienced researchers (see above) also provided advice on the management of such issues.

3.3 Data providers

Please refer to Section E for a summary of the organisations we consulted about provision of data for CALON. As described in Section A we encountered difficulties in communicating effectively with some data providers. However, once we had identified useful contacts, we were able to hold very helpful conversations. Our initial enquiries centred around finding out about the organisations, their processes and the data they were responsible for. For those organisations whose data we proceeded to use, we developed ongoing relationships whereby they continued to provide assistance and advice about the use of their data.

3.4 Information governance specialists

At the early stages of CALON, discussions were held about the nature of the project and relevant ethical requirements. Advice was obtained from the Cardiff & Vale University Health Board's Research & Development department, the South East Wales Ethics Committee, and the Health Research Authority's Confidentiality Advisory Group. Ethical and scientific committees (representing each data provider) reviewed the project protocol and application forms, resulting in further discussions with Cedar in some instances.

3.5 Literature reviewing advice

Cedar's consortium includes staff from Cardiff University's Support Unit for Research Evidence (SURE). Dedicated time was therefore available from an information specialist, who assisted in reviewing the available literature. This provided background information and gave some insight into cardiac ablation procedures, and the history and current status of health data linkage and associated work, both in the UK and elsewhere.

Section E – Choice of datasets

1 Introduction

The suitability of a particular dataset depends on the research question, the type of intervention to be evaluated, how common the intervention is, and which outcomes are of interest. This section will discuss types of datasets, factors to consider when identifying and choosing data sources for research, and present a summary of datasets encountered in the CALON project.

2 Dataset types

When choosing potential sources of data for a study, **be aware of the original purpose for collecting the data, and how they were recorded**.

2.1 Routine clinical data

Routine healthcare data from GP practices and hospitals will have been originally generated to optimise the care provided to individual patients; not for research purposes. Hospital data are usually entered onto electronic health records (EHRs) by individuals trained specifically in clinical coding. These coders are entirely reliant on the information documented in hospital notes by those providing care. In contrast, primary care data are often entered directly by those delivering care, such as GPs, nurses and healthcare assistants. The process of selecting codes is facilitated by the clinical software, which might for example provide 'drop-down' options in response to the text being entered. Additional codes may be added to the available options as a result of local needs. Identification and interpretation of these codes within the context of research can prove challenging.

2.2 Registers

Some datasets have been specifically created for the purpose of research or audit. Bespoke registers are designed to collect data that contributes to the answering of particular research questions, and are therefore usually more focused on specific clinical areas. They may be coordinated by professional societies, and might only exist for a limited length of time. Data are entered by clinicians or administrators acting on their behalf.

When setting up new registers, design the minimum (compulsory) dataset to include identifiers that enable linkage to other datasets. We recommend that early discussions are held with experts from the Farr Institute (for health data) and/or the ADRN (for social science and economic data); see Section C for more information about these organisations. If a register feeds into an appropriate ongoing data linkage repository from the outset, it should facilitate subsequent use of the linked data.

2.2.1 Example of linked registry-based research

The MS (Multiple Sclerosis) Society reported that there were some important research questions yet to be answered about this debilitating neurological condition. Amongst other uncertainties, the number of people affected by the condition and the casemix within the population were not well known. In response to this, a specialist register was set up with links to other sources of data.

The UK MS Register brings together datasets from three main sources:

- Routine clinical data (from HES in England, the Patient Episode Database for Wales (PEDW), and primary care)
- Specialist clinical data collected directly from clinicians in NHS neurology clinics
- Patient-reported data from questionnaires delivered via the internet.

This resource is being used by researchers for various studies, and results have been published about the physical and psychological impact of the condition and how it affects quality of life (Jones et al. 2013a; Jones et al. 2013b).

The platform that has been developed to enable this linkage of different types of information could be repurposed to enable similar types of research. Another example is the "You Tell Us" study being run by Swansea University to learn about patients' views and experiences of a local Health Board. The inclusion of the SAIL team in Cedar's consortium allows us access to these resources and related expertise, and we are keen to take advantage of the opportunities that this relationship offers.

3 Identification of available datasets and selection criteria

As far as we are aware there is not currently a comprehensive list of UK datasets that might be accessed for research purposes, although some resources are listed on the ADLS website. In the CALON project, potential data sources were identified in an ad hoc manner and mainly through:

- Internet searching
- Literature searching
- Conferences
- Word-of-mouth.

Criteria for selection centred on whether the datasets were likely to contain data that could be used to address our research aims (which were to investigate the safety and efficacy of cardiac ablation procedures), and included:

- Geographical coverage
- Temporal coverage
- Completeness
- Data quality
- Linkage potential.

3.1 Geographical coverage

As CALON aimed to contribute to our understanding of cardiac ablation procedures within the UK, we did not consider datasets elsewhere in the world. NICE was particularly keen to include data from England. In addition, we decided to seek data from Wales because we had an established relationship with the SAIL team in Swansea (now part of Cedar's consortium), and to boost total patient numbers. The Welsh contribution is also valuable due to the fact that a much higher proportion of primary care data is available for research. In Wales, 74% of GP practices have signed up to provide data to the SAIL Databank, representing 79% of the Welsh population. The coverage of English sources of primary care data is generally less than 10%.

Scotland is widely recognised as having well-developed data linkage resources. NHS Scotland's Information Services Division holds health-related data for over 5 million people in Scotland, and these data have been used successfully for research purposes. However, significant differences exist between the data available in Scotland and the rest of the UK. For example, NHS numbers in England and Wales are generated by the Personal Demographic Service (PDS), whereas Scotland uses a Central Health Index (CHI) instead of an NHS number. Similarly, Northern Ireland uses another range of numbers to identify patients, the Health and Social Care Number (HSCN).

For these reasons, we opted to limit our pilot project to England and Wales only. The data were similar in both regions (using the same secondary care codes and similar primary care codes), but it was necessary for us to apply for access to data for England and Wales separately. As the project progressed, this had an added advantage (by serendipity rather than design) when linkage of English data was suspended as we were able to complete the work with Welsh data only. See Section A for more information about these issues.

3.2 Temporal coverage

The time periods chosen for retrospective analysis in CALON reflected the history of the cardiac ablation procedures of interest. Very recent data (within the last year or so) were not requested, as data entry may be delayed, and data providers take variable lengths of time to prepare data (such as cleaning and formatting) before they are released. **Those wishing to make use of data soon after they are generated need to ascertain whether this is likely to be possible.** Data providers should be able to supply information about the frequency and timeliness of their releases.

3.3 Completeness

It is unlikely that data are collected from 100% of the population under scrutiny by any one provider. Some collections of records, such as secondary care data from hospitals, would be expected to be fairly complete.

Primary care data have been more difficult to collate. Numerous companies provide software to GP practices for management of EHRs, though a few (such as EMIS, INPS/Vision and TPP/SystmOne) tend to dominate the UK market. Because there are differences in the data that these systems collect, it would be challenging to seamlessly integrate them into one database. Furthermore, many primary care data are only collected on an 'opt-in' basis, and not by default; those compiling datasets for research purposes build up their content through case-by-case recruitment of GP practices. An overview of the coverage of some of the UK primary care datasets is provided in table 2. It is worth noting that most of the sources of English primary care data cover around 10% or less of the population, whereas in the Welsh SAIL Databank permissions have been granted to receive data from 79% of the population (with at least 40% already being available and the remainder expected soon). **Researchers may wish to consider conducting studies using Welsh data if primary care data is of particular interest.**

3.4 Data quality

Data providers often conduct their own analyses of data quality; their **published reports might help inform study design choices**. Missing data are a common problem when making use of routinely collected records. Worth bearing in mind is that data fields tend to be completed more thoroughly when the contents are associated with reimbursements, as the organisations are incentivised by payment (such as those relating to the Quality and Outcomes Framework (QOF)). More information about data specifications and coding issues can be found in Section J.

3.5 Linkage potential

General information about linkage methodology (including deterministic and probabilistic matching) can be found in Section F.

Matching individuals between distinct datasets is greatly facilitated if they contain common unique identifiers (such as NHS numbers). Directly comparable identifiers allow deterministic matching, and relatively high confidence that both records relate to one person. Where such identifiers are not missing, matching may still be possible (on a probabilistic basis) if a combination of other details can be provided, such as name, date of birth, and postcode. **Researchers should confirm the availability and completeness of such identifiers when investigating potential datasets for linkage**.

4 Datasets and linkage repositories

Through the CALON project we have gained some knowledge about the datasets shown in table 2. Other resources might be identified through the ADRN website.

Table 2 Datasets and organisations providing data. Please note that many other datasets are available with potential for research use. See Section C for suggested resources that may assist in identifying datasets of particular relevance to your work.

Dataset	Organisation(s)	Description	Used in CALON?
NICOR Specialist Register – Cardiac Rhythm Management (CRM) dataset	National Institute for Cardiovascular Outcomes Research (NICOR); Healthcare Quality Improvement Partnership (HQIP)	NICOR hosts a collection of clinical data from cardiovascular audits. Data from the Myocardial Ischaemia National Audit Project (MINAP) has been linked and used widely for research purposes. CALON created a new, temporary link between the CRM dataset and routine clinical data in Wales. PROMs data is also stored by NICOR.	Yes, to identify patients who underwent ablation.
Secure Anonymised Information Linkage (SAIL) Databank	SAIL; Health Information Research Unit (HIRU); Centre for Improvement in Population Health through E-records Research (CIPHER)	A data linkage repository that collects data across the whole of Wales. Includes secondary care inpatient data from the Patient Episode Database for Wales (PEDW), primary care, outpatient and death data. 74% of GP practices in Wales have signed up to provide data.	Yes, to obtain outcomes from primary and secondary care data for Wales.
Clinical Practice Research Datalink (CPRD) – Gold dataset	CPRD *formerly General Practice Research Database (GPRD),* Medicines and Healthcare Products Regulatory Agency (MHRA)	Primary care data, covering around 9% of the UK population. CPRD work in partnership with HSCIC to provide linked data (including HES and ONS mortality) from practices that have consented to linkage (about 70% of contributing practices from England). CPRD provides data on a commercial basis to academia and industry.	No. Application was approved but linkage and release of data was suspended due to events at a national level (see Section A).
Hospital Episode Statistics (HES)	Health and Social Care Information Centre (HSCIC) *formerly the NHS Information Centre (NHS IC)*	Secondary care data from England. Includes the majority of NHS hospital inpatient and outpatient records. Some Accident and Emergency data also available.	No. Applied for data via CPRD but not released within project timescale (See Section A).

Dataset	Organisation(s)	Description	Used in CALON?
ONS Mortality	Office for National Statistics (ONS)	Deaths registered in England and Wales. Includes date and cause of death.	No. Requested as linked data via CPRD. Used WDS for Wales.
Welsh Demographic Service (WDS)	NHS Wales Informatics Service (NWIS)	Administrative information (demographic data) for NHS patients in Wales. Includes name, address, date of birth, GP and NHS number.	Yes. Used for matching of patients and for date of death (if applicable).
The Health Improvement Network (THIN)	Cegedim Strategic Data Medical Research UK (CSM MR UK)	A dataset based on primary care data, mainly from England. Contains data from >12 million patients. Contributing practices use Vision software, and overlap with CPRD by approximately 50%. Data from around 180 practices had been linked to HES data by the end of 2014.	No. Limited linkage to HES/ONS at time of enquiry.
ResearchOne	TTP (in collaboration with University of Leeds).	Not-for-profit database for England, including primary and secondary care data. Based on SystmOne GP software. Data are anonymised at source and OpenPseudonymiser software is used to automate linkage.	No. Linkages are ad hoc; all organisations would need to use OpenPseudonymiser.
Work and Pensions Longitudinal Study (WPLS)	Department for Work and Pensions (DWP)	Existing dataset has linked benefit information with employment data (from HMRC) since 2004. May be made available through the Administrative Data Research Network (see Section C).	No. Data were not yet available for linkage by researchers.
CALIBER	Farr Institute @ London incorporating NICOR	Data linkage repository with a focus on cardiovascular disease research. Contains linked data from CPRD, MINAP, HES and ONS mortality.	No. The CALIBER team only wished to add new datasets (such as CRM) on the basis that they could be reused for other research purposes (not possible with CALON).
QResearch	University of Nottingham/EMIS	Not-for-profit primary care database available to academics employed by UK universities for research purposes. Contains records of over 13 million patients, but maximum number of records supplied to researchers is 100,000. Data come from EMIS practice software.	No. Linked data are only available for analysis at the University of Nottingham. Needs ethical approval.

Section F – Linkage methodology

1 Introduction

Methods used to match individual patient records and link large datasets are complex and require specialist skills in informatics. A small number of organisations are able to provide these services in the UK, and data linkage researchers should be able to demonstrate that they have secured the full support of a reputable organisation to assist in this process within a secure environment.

This toolkit predominantly aims to describe the process for obtaining and using data from individually linked records, and we provide a summary of a generic linkage process below. Also discussed is the simpler option whereby an identical variable (such as procedure) is found in two different datasets, allowing comparison of the datasets without linking patient records at an individual level.

CALON

In the CALON project, linkage of Welsh records was conducted by the SAIL team at Swansea University, part of Cedar's consortium. The Health and Social Care Information Centre (HSCIC) had agreed to conduct the linkage of records from English sources, but this part of the project did not proceed due to external delays, as described in Section A. The intended data flow for the project as it was originally designed across England and Wales, and details of the linkage process in Wales, are described in part 4 below.

2 The data linkage process

Data linkage can be used to match records that relate to the same individual person. It can also be used to match records of families, places or events, but for the purposes of this toolkit we are assuming that matching is at person level. The summary presented here is a simplified version of content from the Introductory Analysis of Linked Health Data course (January 2014), hosted by Swansea University and taught by Professor David Preen (University of Western Australia).

2.1 Steps in the process

The data linkage process goes through a number of different steps (illustrated in figure 2):

1. Firstly the data are prepared, quality checked and formatted.

2. Files are blocked together in a way that increases processing efficiency (as these are usually very large datasets). Records are first grouped by one or more identifiers (such as forename, surname, sex and date of birth), so that comparison of records (see step 3) is conducted within these groups; this means that each individual record does not need to be checked against every other record in the entire dataset.

3. Pairs of records are systematically checked against others to determine whether they relate to the same individual. This process is achieved through deterministic or probabilistic matching, or a combination of the two (see below).

4. If a data linkage repository is being created (see Section B), a file of the links used for matching is stored.

5. The matched records are merged using the identified links, (usually) resulting in a single composite record. Some quality checking may also be undertaken at this stage.

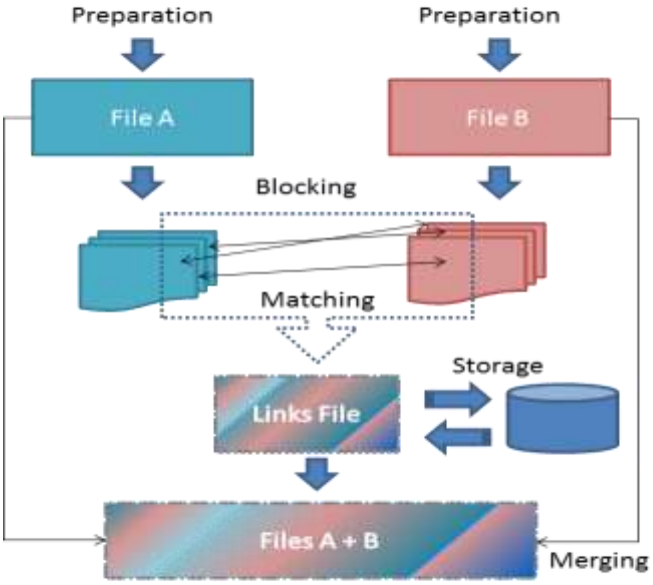


Figure 2 Steps in the data linkage process (adapted from Introductory Analysis of Linked Health Data course, Swansea University and the University of Western Australia)**.**

2.2 Deterministic and probabilistic matching

As a researcher, it may not be necessary to understand the technical details of how the above process achieves its goal, but it is helpful to have some appreciation of the matching stage.

Pairing up two records that are believed to relate to the same individual relies on the type and level of details available within those records. Some fields have a high predictive value in correctly identifying a person, especially where each entry is unique (such as NHS or National Insurance number). Other fields are not unique for each individual, but in combination with other fields may help to identify a person correctly. Examples of such partial identifiers include name, sex, date of birth or postcode.

Deterministic matching is possible if all of the fields being used for the matching process are successfully matched (at least within defined limits). This may be achieved using unique identifiers, or ‘fuzzy’ matching using a combination of partial identifiers that are in agreement between datasets. Where some of the entries vary between datasets but other fields agree, probabilistic methods can be used to weight the probability of these similarities occurring by chance; a threshold is set to decide whether or not to accept the match. The discriminating power of different identifiers can be calculated, based on their sensitivity and specificity. Even where a unique identifier is available, using probabilistic matching in addition yields better results, because there is always an error rate in the unique identifier.

Those conducting linkage services on behalf of researchers should be able to provide information about matching methods used and success rates. This may influence the confidence that is placed in the final study results.

2.3 Trusted third parties

In order to protect confidentiality, methods have been developed whereby demographic details (required for matching records) are not provided to researchers, so that the data are 'pseudonymised'.

In linking data from two providers, a common mechanism is to use a Trusted Third Party (TTP). This organisation is provided with demographic information from both datasets, conducts linkage, and assigns a project-specific identification (ID) code to each record. They are not provided with any clinical data. This method means that no-one other than the original data provider is able to view the identity of an individual alongside their clinical records, whilst allowing researchers to combine information from different sources at an individual person level. Part 4 below shows how this method is being used in the CALON project.

2.4 Data linkage software

Whilst the Cedar researchers do not have experience in using them, we are aware that various data linkage software packages are available, each with its own strengths and weaknesses (Ferrante, Boyd 2012). The main package which we have been informed about is the OpenPseudonymiser, developed by the University of Nottingham and used by ResearchOne.

3 Comparison of datasets without data linkage

Patrick et al. (2012) compared numbers of other interventional procedures carried out in secondary care settings (as recorded in HES) with numbers according to specialist registers. Though the total numbers were 'matched' by hospital and by year, no attempt was made to confirm that the records related to the same individual patients. No further analyses (such as procedural outcomes), were presented in this paper.

It has been demonstrated that this approach may be used to approximate the coverage sensitivity of data sources, if it can be safely assumed that the majority of those 'overlapping ' records did indeed correspond to the same patients. This has the benefit that data can be accessed relatively quickly (especially by those who already have licences to use HES/PEDW). However there are also limitations to the amount of information that can be gleaned using this method. For example, it relies on the same procedure codes being used in both datasets, or at least there being one common procedural code that can be used. Section J provides more information about defining variables and coding systems.

CALON

In CALON, definitions of procedures in the specialist register were considerably different from those that were assigned using standard secondary care procedure codes (OPCS). This offered an advantage in that additional detail (such as ablation energy source) could be provided by the register (since it is not available from routine hospital records). The secondary care records, on the other hand, provide some longer term outcomes and the possibility of identifying unanticipated safety events. Matching of individual patient records allows us to be relatively confident that the observed outcomes are occurring in the same patients who have undergone particular procedures.

Linkage at an individual level is desirable where different elements of the research question are stored in separate datasets. In CALON, only the register held specific details about the procedures, whereas only the routine data provided long-term outcome data.

4 CALON project linkage processes

4.1 Overview of CALON data flow

CALON

The CALON project aimed to link person-level data from a specialist register (provided by NICOR) to routine GP and hospital data in England and Wales. Welsh primary and secondary care records are found in the SAIL Databank, and NWIS acted as their TTP for linkage purposes. In England, CPRD provide primary care records whereas secondary care records (HES) are provided by HSCIC. A different department at HSCIC acts as a TTP and is responsible linkage of English records. All project data was then to be stored in the secure SAIL Databank.

Figure 3 illustrates the planned transfer of data between different organisations in England and Wales. In reality, events beyond our influence (see Section A) meant that it was not possible to include English data within the time allowed for project completion, and so only the Welsh processes were accomplished in this pilot study.

You may wish to refer to our list of abbreviations in the introductory section. Further details about the organisations involved in CALON can be found in Section E.

Full details of the methods that are used to link records and protect the privacy of individuals through use of the SAIL Databank can be found in papers by Ford et al. (2009), Lyons et al. (2009), and Jones et al. (2014).

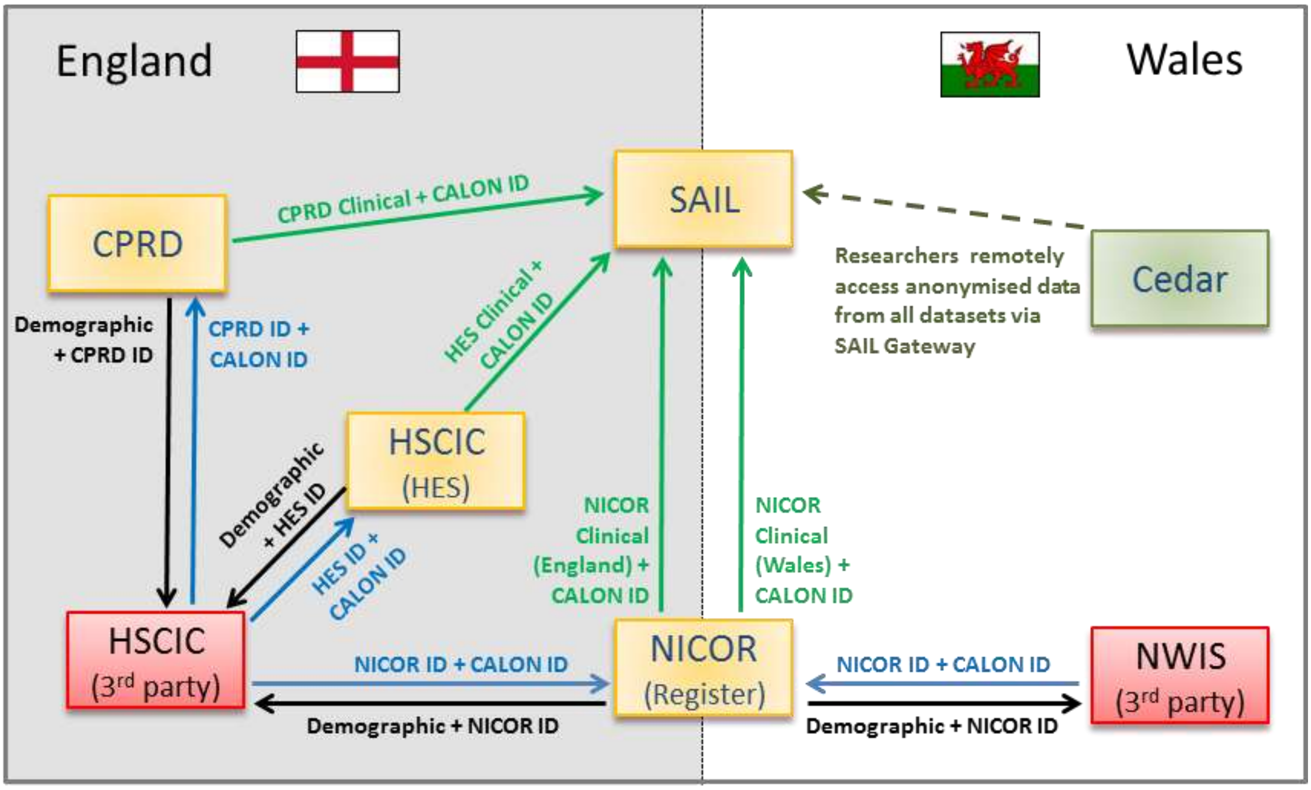


Figure 3 CALON project data flow (as originally planned)

The direction of data flow in figure 3 can be summarised as follows:

1. Demographic data (such as NHS number, name and date of birth) is sent from data provider to TTP with local dataset ID code.
2. TTP links datasets (based on demographic data) and assigns CALON project ID code.
3. TTP returns list of local dataset ID codes with associated CALON ID codes to data provider.
4. Data providers supply clinical data to SAIL with associated CALON ID codes, having removed local dataset ID code and demographic details.
5. Clinical records are combined at SAIL, identified only by CALON ID codes.

There may be additional steps required for some data providers. The actual process for linkage of data between the NICOR registry and routine Welsh data is summarised below. An additional step was included to ensure that NICOR only released clinical data for patients with corresponding records in SAIL. The standard SAIL Anonymous Linking Field (ALF) served the same purpose as a CALON ID code.

4.2 Summary of split-file linkage process for CALON

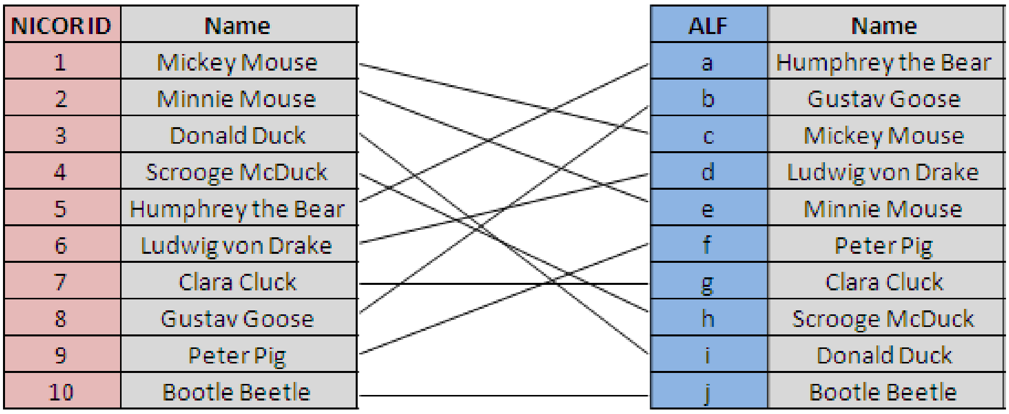
CALON

The following process illustrates the split-file technique used to link the new dataset (from the NICOR register) into the SAIL Databank, with assistance from a TTP (NWIS). Information about the organisations involved can be found in Section E.

1. NICOR sent demographic data to NWIS.

NICOR ID	Name
1	Mickey Mouse
2	Minnie Mouse
3	Donald Duck
4	Scrooge McDuck
5	Humphrey the Bear
6	Ludwig von Drake
7	Clara Cluck
8	Gustav Goose
9	Peter Pig
10	Bootle Beetle

2. NWIS matched the NICOR demographic data to their administrative records (WDS).



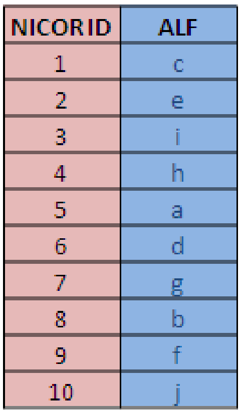
NICOR ID	Name
1	Mickey Mouse
2	Minnie Mouse
3	Donald Duck
4	Scrooge McDuck
5	Humphrey the Bear
6	Ludwig von Drake
7	Clara Cluck
8	Gustav Goose
9	Peter Pig
10	Bootle Beetle

ALF	Name
a	Humphrey the Bear
b	Gustav Goose
c	Mickey Mouse
d	Ludwig von Drake
e	Minnie Mouse
f	Peter Pig
g	Clara Cluck
h	Scrooge McDuck
i	Donald Duck
j	Bootle Beetle

3. NWIS assigned an Anonymised Linking Field (ALF) to each matched record from NICOR.

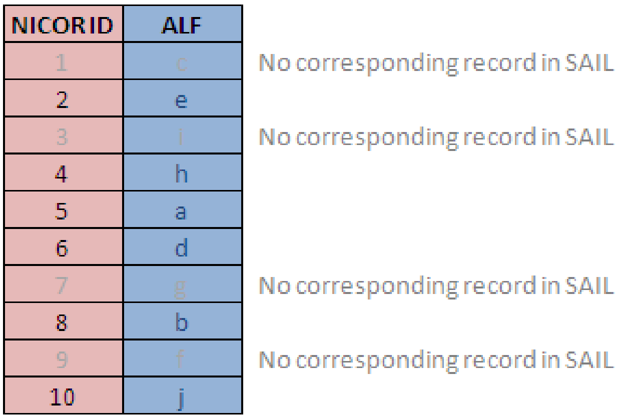
NICOR ID	ALF	Name
1	c	Mickey Mouse
2	e	Minnie Mouse
3	i	Donald Duck
4	h	Scrooge McDuck
5	a	Humphrey the Bear
6	d	Ludwig von Drake
7	g	Clara Cluck
8	b	Gustav Goose
9	f	Peter Pig
10	j	Bootle Beetle

4. NWIS removed demographic data and sent the ALF and NICOR ID to SAIL. The NICOR ID was encrypted before being used by the SAIL analyst.



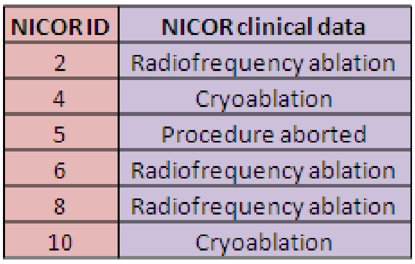
NICOR ID	ALF
1	c
2	e
3	i
4	h
5	a
6	d
7	g
8	b
9	f
10	j

5. SAIL analysts checked the ALF against their clinical (GP and hospital) records.



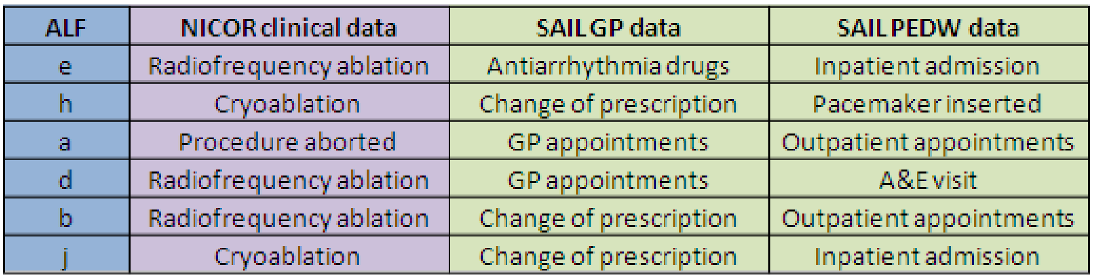
NICOR ID	ALF	
1	c	No corresponding record in SAIL
2	e	
3	i	No corresponding record in SAIL
4	h	
5	a	
6	d	
7	g	No corresponding record in SAIL
8	b	
9	f	No corresponding record in SAIL
10	j	

6. SAIL sent decrypted IDs of matched records to NICOR, who returned the corresponding clinical data from the register.



NICOR ID	NICOR clinical data
2	Radiofrequency ablation
4	Cryoablation
5	Procedure aborted
6	Radiofrequency ablation
8	Radiofrequency ablation
10	Cryoablation

7. SAIL attached the ALF, and used it to link to their GP and hospital (PEDW) records.



ALF	NICOR clinical data	SAIL GP data	SAIL PEDW data
e	Radiofrequency ablation	Antiarrhythmia drugs	Inpatient admission
h	Cryoablation	Change of prescription	Pacemaker inserted
a	Procedure aborted	GP appointments	Outpatient appointments
d	Radiofrequency ablation	GP appointments	A&E visit
b	Radiofrequency ablation	Change of prescription	Outpatient appointments
j	Cryoablation	Change of prescription	Inpatient admission

Section G – Information governance

1 Introduction

Protecting the privacy of patient records is an essential requirement in linking and using individual health and social data, and has received much attention in the UK in recent years. Various safeguards have been introduced including technological, procedural and legislative measures. At the early stages of a project using observational data, it is important to consider the study type and corresponding levels of risk involved, and make sure appropriate measures are in place from the outset.

Whilst we refer here to our experiences with the CALON project, it is important to recognise that **every study is unique and approaches to information governance may differ** depending on the design. We discuss the classification of projects as research (or not), and the impact of this decision on CALON. Data protection legislation and guidance is considered, as well as system level security measures for processing.

2 Is it research?

One of the first questions that should be asked of a study is whether it falls into the category of research or not. If designated as research the study must comply with the Research Governance Framework, or its forthcoming replacement, the UK Policy Framework for Health and Social Care. Compliance can impose additional, and sometimes lengthy, requirements during the study set-up phase. These may include review by a national Research Ethics Committee (REC), involvement of the National Institute for Health Research (NIHR)/National Institute for Social Care and Health Research (NISCHR), and research governance at each NHS organisation. Applications to relevant bodies for research studies must be made through the Integrated Research Application System (IRAS).

Research studies using anonymised data may be eligible for proportionate review by the NHS Health Research Authority (HRA). This is an expedited service for studies which have limited risk and burden to the participant. The proportionate review service aims to provide a decision within 14 days after receipt of a valid application. Social care research (not involving clinical interventions) may be reviewed outside of the NHS context by the National Social Care Research Ethics Committee.

Projects that do not fall into the category of research might include audit, service evaluation, public health surveillance, equipment or system testing, and satisfaction surveys. The distinction between research and non-research is not always clear, and observational studies that make use of existing data may be particularly difficult to classify. In determining whether or not the work would fall into the 'research' category, **we recommend that resources available from the HRA are utilised**. The HRA's website provides a useful 'Decision Tool' and other sources of guidance on this topic. Despite the additional regulatory conditions imposed, it is our observation that the additional scrutiny that research projects undergo can reassure data providers and smooth negotiations later in the process.

It is worthwhile discussing the requirements of data providers at an early stage, to gauge their attitude towards research and ethical approvals. Some data providers, such as SAIL, do not require ethical approvals, as the data are anonymised and all project proposals undergo scrutiny by an independent information governance review panel. On the other hand, other data providers may be reluctant to release their data without the added assurance of a favourable opinion from a research ethics committee. Some organisations explicitly state that they are only able to contribute data towards projects designated as research, whereas others may only support non-research activities (such as audit or service evaluation).

Ultimately the local R&D department will make the decision, but it may be possible to influence their deliberations. R&D departments may be unfamiliar with research using large datasets of linked data, and the appropriate classification may not be obvious. The result may depend on how the project is presented and whether the emphasis is placed on research type activities or not.

2.1 CALON decisions

CALON

CALON was designated as Service Evaluation by the Cardiff & Vale University Health Board's department for Research and Development (R&D). It therefore did not require review by a REC, and also fell outside of the scope of the HRA's Confidentiality Advisory Group (CAG).

The decision not to classify CALON as research impacted on later discussions with data providers. One organisation expressed unease about sharing their data when specific ethical opinion had not been sought. In this instance, further discussion and supplying evidence of data protection measures provided sufficient assurance.

3 Data protection measures

In designing a study, efforts must be made to consider how data will be protected throughout the project, ensuring that legal requirements are adhered to and that the computing infrastructure is secure.

3.1 Legislation and guidance

In Section A, we discussed the proposed EU General Data Protection Regulation and its potential impact upon research using observational data. Until such Regulation comes into force, the Data Protection Act (1998) outlines the methods in which personal confidential data may (and may not) be obtained, processed, stored and used in the UK. It is based on the EU Data Protection Directive (1995).

The Caldicott2 information governance review Information: to share or not to share? (2013) notes that "the complexity, confusion and lack of consistency in the interpretation of legal and governance requirements can sometimes hamper research", and that data providers tend to err on the side of caution. It is also observed that the terms used to describe types of data (such as 'personal' or 'identifiable') are inconsistently defined. The report proposes a simple framework based upon three forms of data:

i. De-identified data for publication

ii. Personal confidential data

iii. De-identified data for limited disclosure or access

Release of de-identified data for limited disclosure or access should be safeguarded by the requirement for a contractual agreement and conformance to data stewardship functions (such as those seen in data sharing agreements, see Section H). Irreversibly anonymised data are not covered by the Data Protection Act (Boyd, 2003).

The Health Research Authority's Confidentiality Advisory Group (HRA CAG) makes provision for the common law duty of confidentiality to be set aside, allowing access to patient information (without requiring consent) under certain conditions in accordance with Section 251 of the NHS Act (2006). Al

The HSCIC has also recently published a Code of Practice on Confidential Information for England (HSCIC, 2014).

3.2 Data processing

Common requirements laid out by data providers in applications and data sharing agreements are listed in Section H. These requirements include the provision of assurance that information systems for processing and storage of linked data are secure.

3.2.1 IT System requirements

Various methods are used to link and access data (see Section F), and likewise there are different ways in which computer systems have been designed to protect data.

CALON

In CALON we were fortunate to have support from our consortium partners at SAIL. They agreed to receive and store all project data within the SAIL Databank, which is included in the ADRN's list of UK Safe Centres. Information about measures used to protect the privacy of individuals whose data are stored in the SAIL Databank can be found in articles by Ford et al. (2009) and Jones et al. (2014).

Researchers should ensure that methods used to link and store project data are robust. Whilst some degree of protection is afforded for those utilising NHS IT systems, this alone may not be sufficient to satisfy the requirements of data providers. **All data should be linked and processed within a secure environment, such as the SAIL Databank**.

3.3 Safe researcher training

Some data providers require completion of an approved course by individual researchers before they will permit access to their data. The Administrative Data Liaison Service provides training in the safe and responsible use of administrative data. The Safe Researcher course has been endorsed by major UK data providers and the Information Commissioner's Office. A similar course is available for completion online from the ScottisH Informatics Programme.

Section H – Applications to data providers

1 Introduction

Each data provider will have its own application processes and requirements to check the scientific validity and information governance measures of a proposed project before releasing data to researchers. This section provides an overview of some issues to consider when making applications, listing common requirements of data providers. We conclude by describing our experience in submitting applications for the CALON project, with an example of difficulties faced.

2 Application processes

As discussed in Section A, one of the main challenges in dealing with applications to data providers is the wide variation in their processes and requirements. Improvements are expected in the future, as organisations move towards more collaborative arrangements. Facilitating this transition is one of the responsibilities of the Farr Institute (see Section C). We have been informed that they are aiming to provide data from multiple sources upon receipt of a *single* application; this development should greatly reduce the administrative burden on researchers at the early stages of complex projects, and is eagerly anticipated. This has already been implemented at a national level in Wales by SAIL; it is hoped that in future the co-ordination of efforts will be extended throughout the UK.

In the meantime, applications must be submitted independently to several organisations (though some, such as CPRD and HSCIC, already work in partnership under single applications). Our suggested approach is to **contact organisations early for details of their particular processes**. Some may request that application forms are first completed and submitted as an *informal* enquiry, for review and discussion prior to a more formal request for data (using the same or another application form).

2.1 SAIL application process

Some processes are already clearly documented and accessible, such as those found within the SAIL Data Management Policy; a flow diagram of the user journey can be viewed in figure 4. Our experience was that some organisations were less forthcoming with this information, preferring to provide less formal guidance based on individual projects. In particular, the timescales for completing each stage of the process are likely to vary by organisation and over time (depending on their workload and demand).

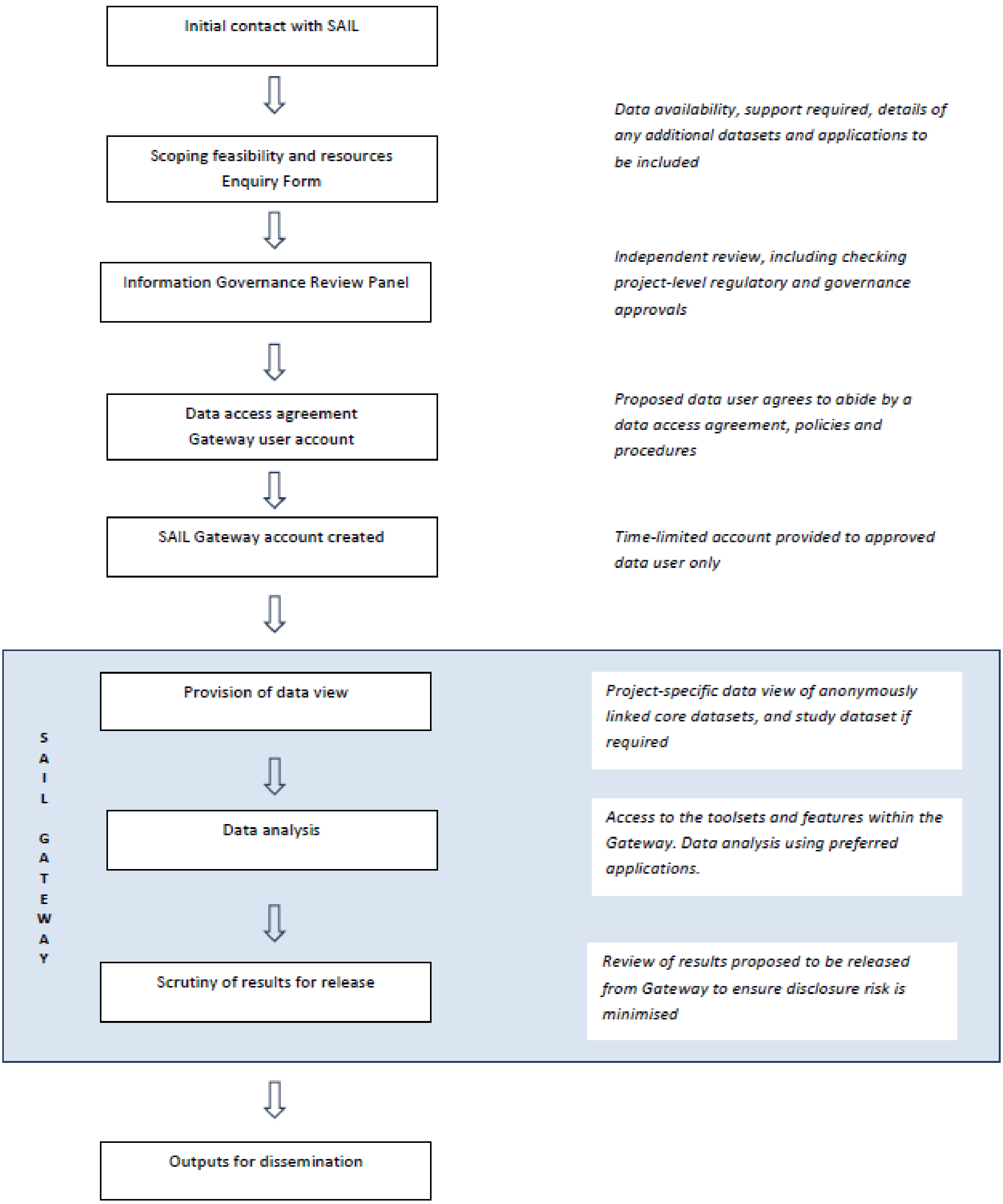

Initial contact with SAIL
Scoping feasibility and resources Enquiry Form
Data availability, support required, details of any additional datasets and applications to be included
Information Governance Review Panel
Independent review, including checking project-level regulatory and governance approvals
Data access agreement Gateway user account
Proposed data user agrees to abide by a data access agreement, policies and procedures
SAIL Gateway account created
Time-limited account provided to approved data user only
SAIL GATEWAY
Provision of data view
Project-specific data view of anonymously linked core datasets, and study dataset if required
Data analysis
Access to the toolsets and features within the Gateway. Data analysis using preferred applications.
Scrutiny of results for release
Review of results proposed to be released from Gateway to ensure disclosure risk is minimised
Outputs for dissemination

Figure 4 SAIL data user journey (from SAIL Data Management Policy v1.1)

3 Data providers’ requirements

3.1 Application form

Requests for data are usually made by completing an application form, available from the data provider. The contents of these forms vary considerably; here we list some of the typical features:

1. Research team
    - Principal Investigator
    - Contact details
    - Team roles
        - Who will access/analyse project data?
2. Requesting organisation
    - Academia/industry/NHS?
3. Brief description of project
    - Purpose/aims
    - Analysis plan
    - Expected outputs
    - Anticipated impact
    - Lay summary
    - Project timescales
4. Data extract request
    - Date range
    - Geographical coverage
    - Any patient level/identifiable details?
    - Variables/codes required
    - Regular data updates required?
    - Data linkage required?
        - If so, datasets to be linked
5. Will additional information be sought from patients or clinicians (for example through questionnaires)?
    - Contact and consent plans
6. Have project plans previously undergone ethical or scientific review?
    - Outcome(s)
7. Research team experience
    - Statistical analysis
    - Use of large datasets
    - Publications
8. Source of funding
9. Level of support requested
    - Data access only
    - Case definition
    - Statistical advice
    - Full analysis and production of report.

3.2 Supporting documents

In addition to a completed application form, **it may be necessary to provide other documents to support the application**. This could include a detailed project protocol (see Section I).

Other documentation may be requested to provide assurance of information governance measures. Descriptions or evidence of data management practices might include:

- Where data will be stored
- How data will be accessed (and by whom)
- IT systems and security measures in place (see Section G)
- Retention period/data destruction plans upon completion.

Associated data protection, confidentiality and data usage policies may be requested. A signed Data Sharing Agreement is commonly required.

3.2.1 Data sharing agreements

Data sharing agreements (DSAs) are effectively the terms and conditions applied by data providers to the use of their data. These stipulations often go beyond basic legal requirements. By signing a DSA, each party agrees to adhere to the principles and procedures as defined by the document. **Applicants should carefully read these terms before signing**.

It may be appropriate to negotiate amendments to the text; depending on the context and interests of the respective parties, these discussions may be complex and time-consuming. Incorporation of a new dataset into a data linkage repository for ongoing use (see Section B) is likely to involve some very careful negotiations. Anecdotally we have been told that **reaching agreement may take several years** in such circumstances. Some organisations are very cautious due to negative experiences in the past. It seems that misaligned organisational aims and previous abuses of trust may hinder the process of reaching consensus about how the data are used. **Demonstrating prior responsible use of data is likely to strengthen the position of data applicants**.

Constraints that are commonly applied by data providers include:

- Data fields requested must be justified by the scientific design of the study, and only used for the purposes specified (in the application form or protocol)
- Subsequent amendments to the study design or protocol must be authorised
- Data are to be stored in a secure environment that meets specific standards
- Access to data is limited to named individuals only
- Data cannot be shared with third parties without the express permission of the original data provider
- Data must be returned or destroyed upon completion of the project, within a specified timescale and by approved methods
- The use of the dataset and/or contributions of organisations must be acknowledged in publications
- The data provider must receive notification or copies of draft outputs ahead of publishing and/or presenting results (up to one month in advance of submission)
- Individuals must remain anonymous, minimising the risk of statistical disclosure.

4 CALON applications – example of problems encountered

In Section A we noted our difficulties in establishing and maintaining effective communications with data providers. These miscommunications impacted to some extent also on our applications for data, illustrated by the following example.

CALON

Here we refer to two organisations – the data provider (or 'custodian'), and the organisation that originally commissioned collection of these data. Several people are also mentioned; these all worked for the data provider, and each letter (A-F) refers to a different individual.

1. In June 2013 we made initial contact with a data provider by email, with a very brief description of the project and our intentions. We received a response from one of their representatives (A). They suggested that we submit a completed an application form as an *informal* application, as their *research group* was due to meet in July 2013.
2. The informal application was submitted to this data provider in early July 2013 as advised. We deliberately omitted some details, which were intended to be added into the formal application at a later date.
3. In August 2013, Cedar received a response (from 'B'), who said that the *research group* was interested in supporting this work, and that the application had been forwarded to an *academic group* for review. We were not advised to amend or update our application, and so awaited further instruction.
4. In October 2013, 'C' informed us that we also needed to apply to the original commissioners of the data collection that we wanted to access. This involved completion of a different application form and provision of various supporting documents.
5. The application to the commissioners was submitted, via the data provider as instructed, in November 2013.
6. Later in November 2013, the data provider's *research executive* and *academic group* held a discussion with NICE about the project; various criticisms were raised about the original application form that had been submitted. When Cedar was later informed about this, it became apparent that the initial *informal* application had been considered as a final, formal request for data.
7. Shortly after this meeting, approval for release of data was granted by 'D', but under a number of additional conditions. At this stage a named contact was provided (E) to coordinate project communications on behalf of the data provider; this was a helpful decision, but **would have been much more helpful if an individual had been nominated to deal with the project from the outset**.
8. In January 2014 it became evident that our application to the commissioners (from November 2013) had not been forwarded by the data provider. This application was therefore re-submitted.
9. In February 2014 we were advised (by 'F') that the commissioners had also agreed to release of data. Written confirmation of this was provided on request.

As indicated by the dates on the previous page, the entire application process (for one dataset) took eight months, in which time other elements of the project were prevented from progressing. Although the data provider in this example did provide a flow chart of its study approval processes, our experience was that their own staff did not follow the specified procedure.

The main obstacles appeared to occur due to the involvement of numerous different organisational representatives, with no clear and consistent internal communication trail. The application appears to have been scrutinised by three different committees (all representing the data provider); it was not clear what the functions of all of these groups were or why there appeared to be duplication of effort.

It should be noted that these organisations were not unusual in these weaknesses; we also encountered similar difficulties with another unrelated data provider.

Cedar had limited influence on other organisations' in-house processes. Some of the lessons learned from our experiences with application processes were:

- **Make sure that any informal applications are clearly marked as such.**
- **Keep thorough records of communications, including names of contacts, dates and advice received or decisions made.**
- **Periodically share a brief summary of relevant project history with the data provider, to increase their awareness of progress to date.**
- **Make every effort to identify a helpful individual (within each organisation) who might coordinate and facilitate communications. Keep them informed of any project developments or difficulties relating to their organisation.**

See also Section A, for more comments about communicating with external contacts.

Section I – The project protocol

1 Introduction

The project protocol is a document in which the plans for the project are formally recorded. This section provides guidance about when to write a protocol, why it is needed, what information to include, and protocol amendments.

2 Development

Some (though not all) data providers will want to see a project protocol before they agree to release their data. If a protocol has not been fully developed, beforehand, this work may be conducted in parallel with completion of application forms (see Section H for more about applying for data). **We recommend that the requirements of data providers are checked before the protocol is written**, as they may dictate the format and expected content.

Data providers may be able to provide advice whilst the protocol is being developed, perhaps even giving feedback on a draft version. Such **a collaborative approach should be sought and taken advantage of whenever possible**, as it is in the interest of both the researchers and the data providers that the best use is made of the data.

3 Purpose

Even if a protocol is not requested by data providers, producing one is likely to benefit the project greatly. Summarising all the project plans within one protocol will help to define and clarify what the project hopes to achieve, and the methods that will be used. The project team and steering group (see Section D) should all be involved in contributing to its development; ideally written confirmation of their acceptance of the plans will be obtained.

Having the proposals in writing in advance of receipt of data enhances the scientific integrity of the study, as a subsequent deviation from the plan might legitimately be questioned (see below for information about amendments).

4 Protocol contents

Unlike standard clinical trials, very little general guidance is currently available to direct the content of protocols for this type of observational study.

Recently, efforts have been made to develop formal reporting guidance and increase the transparency of methods used in observational studies, in the form of the RECORD (Reporting of studies Conducted using Observational Routinely-collected Data) guidelines (Langan et al. 2013). Whilst these guidelines relate more to the information that is published upon study completion, reviewing them may influence the content of the project protocol. At the time of preparing this report the RECORD guidelines had not yet been published, but plans are in place to submit the checklist for publication by March 2015.

The Clinical Practice Research Datalink (CPRD) describes its specific requirements for protocol contents on their website, with guidance available to download in PDF format. CPRD also includes a helpful checklist of protocol contents, under the following headings:

- Lay summary
- Background
- Objective, specific aims and rationale
- Study type (descriptive, hypothesis generating, hypothesis testing)
- Study design
- Sample size/power calculation
- Study population
- Selection of comparison group(s) or controls
- Exposures, outcomes and covariates
- Use of linked data (if applicable)
- Data/statistical analysis plan (including plans for addressing confounding and missing data)
- Patient/user group involvement
- Limitations of the study design, data sources and analytic methods
- Plans for disseminating and communicating study results.

See Section J for more information about the process of defining populations, interventions and outcomes, and the challenges of making use of existing codes.

4.1 CALON protocol

CALON

In producing a protocol for the CALON project we found that CPRD were the only organisation we were working with that provided detailed instructions about what to include in the protocol. Our experience was that it was difficult to provide the level of detail requested by CPRD, whilst still adhering to their requirement to 'be succinct' and aim for 5-10 pages of text (on A4 paper). We found that it was necessary to produce several appendices for supplementary information.

We were fortunate that other data providers were less stringent with their expectations for the project protocol, as otherwise there may have been conflicting requirements. The production of multiple versions of a protocol (to suit each data provider) would introduce a high degree of risk, potentially resulting in confusion and even unauthorised use of data.

We now note that CPRD have added further options on their website that allow for a little more flexibility in balancing the need for information against the limitations on protocol length. They also now encourage more of a collaborative approach in designing the overall study; unfortunately this help was offered to us too late in the process for it to be of benefit. In future we would seek such an arrangement wherever possible.

Should researchers encounter inconsistent and conflicting requirements, it is suggested that discussions are held with the respective data providers in order to reach a compromise, allowing for the production of one project protocol only. It is hoped that in the future these difficulties can be avoided as organisations work together more efficiently, integrating their services so that data are provided from multiple datasets in response to the submission of a single application. Co-ordination of these efforts falls within the remit of the Farr Institute.

5 Protocol/study amendments

As with clinical trial protocols, applications for data and associated protocols should be adhered to once they have been approved. If changes need to be made for any reason, data providers may need to be notified. In some circumstances, amendments may need to undergo additional review by ethical or scientific committees. This can be a time-consuming process if multiple data providers are involved. **It is advisable therefore to refine project plans as much as possible before submission of applications (and protocols)** to data providers, to avoid potentially delaying the project at a later date.

In reality, however much planning a project undergoes, there may still be situations beyond the control of the researchers that can impact on the study once it is in progress (see Section A, for an example of this). Information about what constitutes a minor or major change, and actions that should be taken if circumstances impact upon the study design, should be available from data providers. This may be included within a data sharing agreement or similar document (see Section H).

When making applications to data providers, it is often necessary to stipulate the retention period of the study data and the plans for its destruction/disposal upon study completion. If the approved project subsequently experiences unexpected delays and the completion date is postponed, researchers may need to contact individual data providers to obtain permission to retain data for a longer period. Again, this may necessitate contacting multiple organisations.

Care must also be taken if initial applications are submitted to different data providers at different times, as project plans may have changed between submissions; there is potential that an organisation might not have approved the most recent study design. Again, the proposed streamlining of applications (facilitated by the Farr Institute) should mitigate this risk, and similarly should also allow amendments to be managed through a standardised route.

Section J – Data specifications

1 Introduction

Unlike studies in which data are prospectively collected, such as clinical trials, the majority of projects that make use of administrative datasets for observational research are entirely reliant on analysis of retrospective, routinely collected data. Whilst there are benefits in using such 'real-world' data (see Section B), this study design introduces other complexities, of which one of the main challenges is appropriate use of codes.

In this section we consider:

- Implications of retrospectively accessing data
- Characteristics of routinely-collected data, quality and linkage success
- Classification systems and codes
- Considerations in defining a cohort, interventions, outcomes and covariates.

Some parts of this section will only be directly applicable to studies based on data from the UK, though the broad principles may be generalisable to a wider range of geographies.

2 Retrospective use of data

Projects which make use of routinely collected data will often include design elements commonly seen in other types of research. The population of interest, intervention or exposure, outcome measures and covariates will need to be defined with reference to the research question and any associated hypotheses. The difference is that prospective studies usually permit a much greater control over data collection methods and activities, whereas retrospective analyses are often reliant exclusively on data that have already been collected. Alternatively, it is possible to design a study to prospectively collect data via routine sources, in which case there may be a greater potential to influence the nature and quality of data collected.

When accessing historic data, **it should be recognised that the classification systems** (see below) **may not be the same as those used in current practice**. Finding out the implementation date of each system or version can be helpful in understanding which codes to search for, bearing in mind that there will often be a transition period in which both systems may have been in use concurrently. Similarly, some codes may currently be discontinued, but will still appear in retrospective data views. A common example of this type of event is seen when a drug has been withdrawn from the market.

3 Characteristics of routine data

The data contained within existing datasets may have originally been collected for a number of different reasons, such as for clinical patient management activities, calculating payment for services provided, or for audit purposes. Section E refers to the influence that this might have on the way that these data are recorded, and likewise the impact upon subsequent interpretation and use of results.

3.1 Data quality

A common problem encountered when analysing routine data is missing data, especially in fields that are not required as part of a minimum (compulsory) dataset. Errors in processing of datasets after data entry can also affect the available data, for example if a particular period of time is incompletely transferred. Data can be omitted from any variable; if demographic data are missing, this might impact on matching success if data linkage is being conducted. Depending upon the proportion of data missing, and the importance of those particular fields for the study results, researchers may need to employ a strategy for dealing with these missing data. Statisticians who are familiar with analysis of large datasets should be able to provide appropriate advice.

As well as missing data, there may be other inaccuracies in the data, such as misclassifications. **Researchers should be aware of the potential for errors in a dataset, and mitigate the impact where possible**. Health informatics experts should be able to assist with helping to identify where mistakes have occurred and in rectifying some of these issues. For example, if an entry has been made in a patient record that indicates a contact with clinical services but is dated *after* their date of death, it is very likely that one of the entries was incorrect. The data would then be examined for other clues to verify the situation. It is possible to automate some of the processes that identify this type of error, which is an important consideration when scrutinising very large datasets.

Data providers will generally have their own methods for checking and validating data quality prior to release, but are unlikely to have corrected all errors prior to release of a data extract. **It may be worthwhile obtaining their data quality reports**, where available.

3.2 Linkage matching success

It is possible to estimate the success rate of matching individual patient records between two datasets as part of the data linkage process. Linkage success is dependent on the quality of the identifiers used for matching. Deterministic methods generally produce a higher match rate than is achieved using probabilistic techniques (see Section F). The organisations that carry out linkage services should be able to provide details of matching success rates, which may affect the level of confidence placed in the wider study results.

CALON

Whilst the matching processes were being carried out for CALON, some problems were identified with the records received. Whilst the particular reasons for these errors are likely to have been unique to this study, it did raise our awareness of the difficulty in resolving discrepancies once data have been anonymised, as no one organisation is in possession of the fully linked dataset as well as the patient identifiers.

4 Classification systems and codes

A number of different classification systems and corresponding codes exist within healthcare. Clinical coding specialists undertake dedicated training programmes to learn the skills required to allocate appropriate codes. Bespoke datasets may use their own categories and codes. In selecting fields and codes for research or evaluation purposes it may be beneficial to consult relevant experts in these systems.

For the CALON project we consulted the Classification Standards Manager and members of the Clinical Classifications team at NWIS. We also spoke to dataset managers and clinicians involved in data entry, depending on the dataset of interest.

4.1 Data dictionaries, sample datasets and feasibility testing

For some datasets, it is possible to obtain one or more data dictionaries. These may be documents or electronic databases that list and define the fields and codes associated with that particular dataset. **Obtaining data dictionaries** (where available) **early in the project design stage**, allows researchers to gain insight into the available fields and the granularity of the data, although will not usually indicate how well the fields have been completed. Some data providers may want to see complete lists of data items being requested at the application stage (see Section H for information about applications to data providers).

Data providers may also be able to supply a sample dataset of anonymised records. Whilst these factitious datasets cannot be used for research themselves, it can be helpful to view the typical format and content of the available data. **We recommend enquiring whether this type of resource is available**, as it may not be widely publicised by the data provider.

Data providers will sometimes conduct a feasibility exercise to assist researchers before they submit a full application for data. This might be used to investigate the sample size available within a dataset against specific inclusion and exclusion criteria. It might also indicate potential problems with the planned study design, and so help to direct a more appropriate request.

4.2 Primary care codes

Read codes are the standard clinical terminology system used in UK primary care. As described in Section E, these codes are often added to electronic health records in GP practices directly by healthcare practitioners or sometimes administrative staff. Those responsible for data entry in this context rarely receive substantial training in coding practices. Clinical software systems are designed to facilitate the process, with the incorporation of templates and option lists that may be tailored to suit a particular clinician or practice. Supplementary codes can be added to the systems to address local needs. It should be noted that multiple companies produce differing clinical software systems, which may have some impact upon how Read codes are assigned to patient records.

There have been several versions of Read codes. Version 3 is also known as Clinical Terms Version 3, or CTV3; and has lost the 'Read' name. Read version 2 is also still in active clinical use, and both versions 2 and 3 are amended regularly. A more recent development is the SNOMED CT classification system (see below); adoption of this clinical terminology by primary care systems is expected by the end of December 2016 (National Information Board, 2014).

4.3 Secondary care codes

The main classification systems found in secondary care records are ICD (diagnoses) and OPCS (operations/procedures). The International Classification of Diseases (ICD) is used to classify diseases and other health problems. It is currently in its tenth version (ICD-10), and the ICD-11 revision is due for release in 2017. OPCS refers to the Office of Population Censuses and Survey Classification of Interventions and Procedures, which was formerly known as the OPCS Classification of Surgical Operations and Procedures. The current version is OPCS-4.7.

In HES and the Welsh equivalent (PEDW), each row of data represents a 'finished consultant episode', which covers the length of time that a patient is under the care of one consultant. Patient records can therefore consist of multiple rows of data, even within a single admission or 'spell' (if transferred between consultants). Up to 14 ICD codes and up to 12 OPCS codes may currently be entered onto a patient record per finished consultant episode.

Providers of secondary care employ teams of clinical coding specialists to enter ICD, OPCS and other codes onto electronic patient records. Coders refer to patient notes and other relevant documents (such as discharge letters) when selecting appropriate codes. They are therefore entirely reliant on the quality and legibility of the information recorded by clinicians in existing documentation, and are trained not to make their own conjectures. This has implications for secondary users of these data (such as researchers), as the desired level of detail may not be available.

4.3.1 Safety outcomes

In the CALON project, we used ICD-10 and OPCS-4 codes to define a number of safety events that were likely to be seen after cardiac ablation procedures according to the literature. Our discussions with NWIS revealed that there three main ways in which clinical coders might record procedural complications using ICD-10:

1. Codes that begin with 'T' specifically indicate complications
2. After first entering a code that refers to a condition (such as bleeding), an external cause code is added to indicate that the condition was caused by a procedure
3. Within a chapter relating to a body system, there may be codes that refer to complications that are associated with that particular system (for example, a code for postpartum haemorrhage is found in the chapter for diseases of the circulatory system).

When provided with a list of common safety concerns relating to cardiac ablation procedures, NWIS informed us of the corresponding codes that would be assigned to each condition in the clinical notes (and hence the electronic patient record).

One lesson that we learned here is that it is not always possible to use these terms in 'reverse' in the context of research. For example, there is no specific code for an atrio-oesophageal fistula. When referring to the patient notes, clinical coders would therefore use a generic code (in this case, T81.7 "Other complications of procedures, not elsewhere classified") to indicate that there had been a procedural complication. If this same code were then used from a research perspective to try to quantify the number of atrio-oesophageal fistulae experienced by patients following an ablation, then many other procedural complications would be included in that count. It was therefore not possible to report the incidence of post-ablation atrio-oesophageal fistulae for CALON; instead any record of T81.7 was included in the count for 'Other complications'.

Similarly, the code I63- had been listed alongside both stroke and silent cerebral embolism. Had we counted those conditions separately, the data might have indicated multiple events and led to over-estimation of these complications.

It is important to examine full definition of codes, and consider how they are intended to be used within a study. Try to minimise ambiguity and be aware of the potential unintentional misuse of codes or classifications.

4.4 SNOMED CT

SNOMED CT (Systemized Nomenclature of Medicines – Clinical Terms) is based upon a convergence of SNOMED RT (Reference Terminology) and CTV3. It has been approved as the Fundamental Standard for Clinical Terminology within the NHS in England. SNOMED CT was designed to provide a standardised technology for use across a range of NHS IT systems, including those used in primary and secondary care settings. Terms are available to describe prescribing, referrals, hospital discharges and business processes, and are arranged in interrelating, multilevel hierarchies (White J, Carolan-Rees G 2013). Adoption of SNOMED CT throughout the whole health system is anticipated by April 2020 (National Information Board, 2014).

As there is a lag time between availability and implementation of new terminologies, SNOMED terms may not be found when retrospectively viewing routine datasets. Whilst they therefore currently have limited application at present, their comprehensive coverage of healthcare-related terminologies may prove valuable in the future as the recording of these terms becomes more widespread.

4.5 Codes used in other datasets

Other datasets may make use of recognised clinical classifications, but alternatively might devise their own classifications and codes (or adapt existing ones), to suit their particular needs.

CALON

In CALON we made use of the NICOR electrophysiology/ablation dataset. This dataset had been tailored to collect bespoke codes within fields to facilitate regular auditing by the British Heart Rhythm Society. As this was the case, it was necessary for us to consult with experts who had considerable experience of developing and using these particular data, in order to understand how the definitions were normally used.

5 Definitions

Although codes contribute to the definitions of each study parameter, it is often the case that multiple codes must be used in combination to do so; a single code may not describe the required specification adequately in isolation. In designing a study, operational definitions should be developed and recorded to identify interventions, outcomes and similar elements of the study. The level of detail should permit others to interpret and use these definitions correctly.

5.1 Defining the cohort

Unlike prospective recruitment of patients into a trial, the inclusion and exclusion criteria applied to retrospective studies must be based on data that are already available in existing records. As the study does not interfere with normal clinical care, costs per subject are much lower than those of trials, and sub-group analyses can be performed to account for demographic differences or covariates. This means that data from very large numbers of individuals can be included.

Geography is of relevance, and will depend upon the coverage of the datasets being used, and any regional differences that are likely to exist in data. It is common to incorporate some fields in a study design that relate to geography, such as hospital identifiers or socioeconomic measures (the Townsend Deprivation Index being one example).

A key consideration in defining a cohort will be time. In retrospective analysis of routine data, individuals might have entered and/or left the cohort at any time. Reasons for these events include new diagnoses, procedures, geographical movements, births and deaths. Resolution of a clinical condition might also affect eligibility for inclusion, but is often difficult to ascertain based only on routine data as this would rarely be recorded. Methods have been developed by health informaticians to handle differing lengths of follow-up, the influence of important events that occurred outside of the study period, and transfers between healthcare providers. **We recommend consulting those with experience in such techniques when designing the study**. A basic introduction to these concepts is also provided in the courses run by Swansea University with the University of Western Australia (see table 1 in section C for information about training courses).

One suggestion is to draw a ‘time traveller trace’ to visualise potential scenarios. A fictional example is provided in figure 5. Each horizontal line represents the time that records were available for an individual; other symbols indicate events such as procedures, hospital admissions and loss to follow-up.

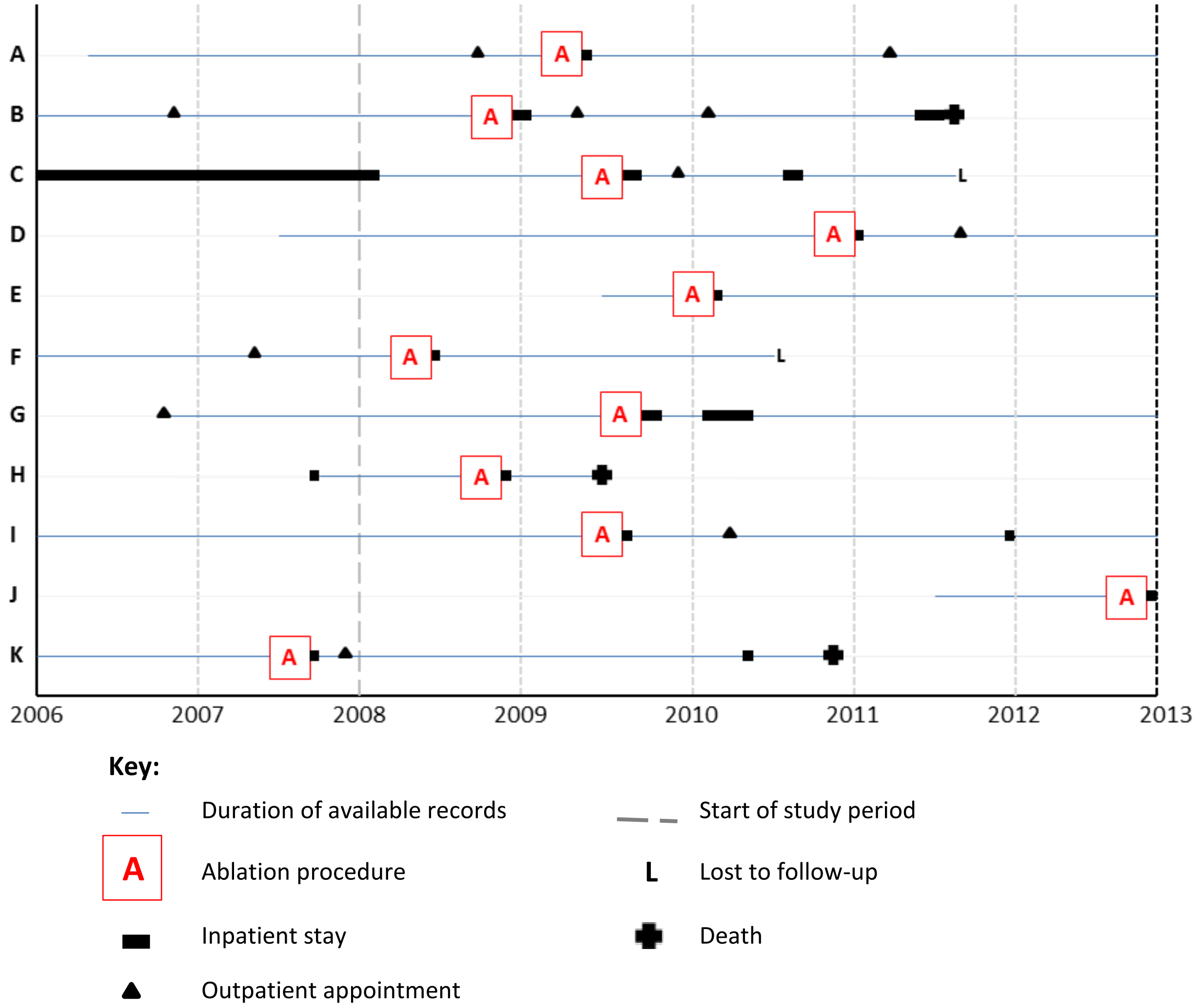


Figure 5 Example of a time traveller trace.

5.2 Defining the intervention and comparators

Not all studies will evaluate the impact of an intervention (such as a procedure), but for those that do, it is important to define the intervention as well as possible. If it has its own code that provides an adequate description whilst being sufficiently specific, then this process may be relatively straightforward. However it can be challenging to define an intervention that relies on a combination of multiple codes, or where codes are ambiguous (Patrick et al. 2012).

CALON

Whilst designing the CALON project we experienced difficulties in defining procedures, both when using routine hospital data and also with the specialist register. In the routine hospital data (HES) we searched for OPCS/ICD code combinations as recommended by NICE, but found much lower numbers of procedures than anticipated. This suggested that either clinicians were not documenting the details we required, or that coders were using other codes/combinations.

Having seen a list of fields and codes used in the specialist register (in the form of a data dictionary), we had expected it to be relatively straightforward to categorise procedures. However, discussions with clinicians (who were involved in using the register) revealed some insights that affected these definitions. The register had not been specifically designed for the purposes of CALON, and it became evident that the data fields and codes were not set up to adequately describe all the procedures of interest to us.

Whilst designing the study we had developed an algorithm in an attempt to differentiate procedure types. We present it here (figure 6) to illustrate the complexity of this task, but would advise against reusing this diagram in other projects. In fact, once we later received the project data, we discovered that it was not possible to use this algorithm to differentiate between procedures in such a way that there were sufficient numbers for statistical comparison.

The expectation that the procedure type could be determined using the NICOR register was the main argument for linking it to the routine data. Once we concluded that the data we had received was not suitable for this purpose, we realised that in fact the outcomes of the main efficacy and safety analyses could have been determined using the routine data alone. In these particular circumstances, the linkage had not added much value to the study, yet had been costly in time, effort and resources to set up. With hindsight it would have been beneficial to run an initial feasibility study based on existing linked data within the SAIL Databank, before deciding whether to proceed with linkage to the register.

The main lessons learned in this process were to **carefully check that any parameters that are key to the success of a project are adequately defined by the available data**, and to **run a feasibility study using existing resources wherever possible**.

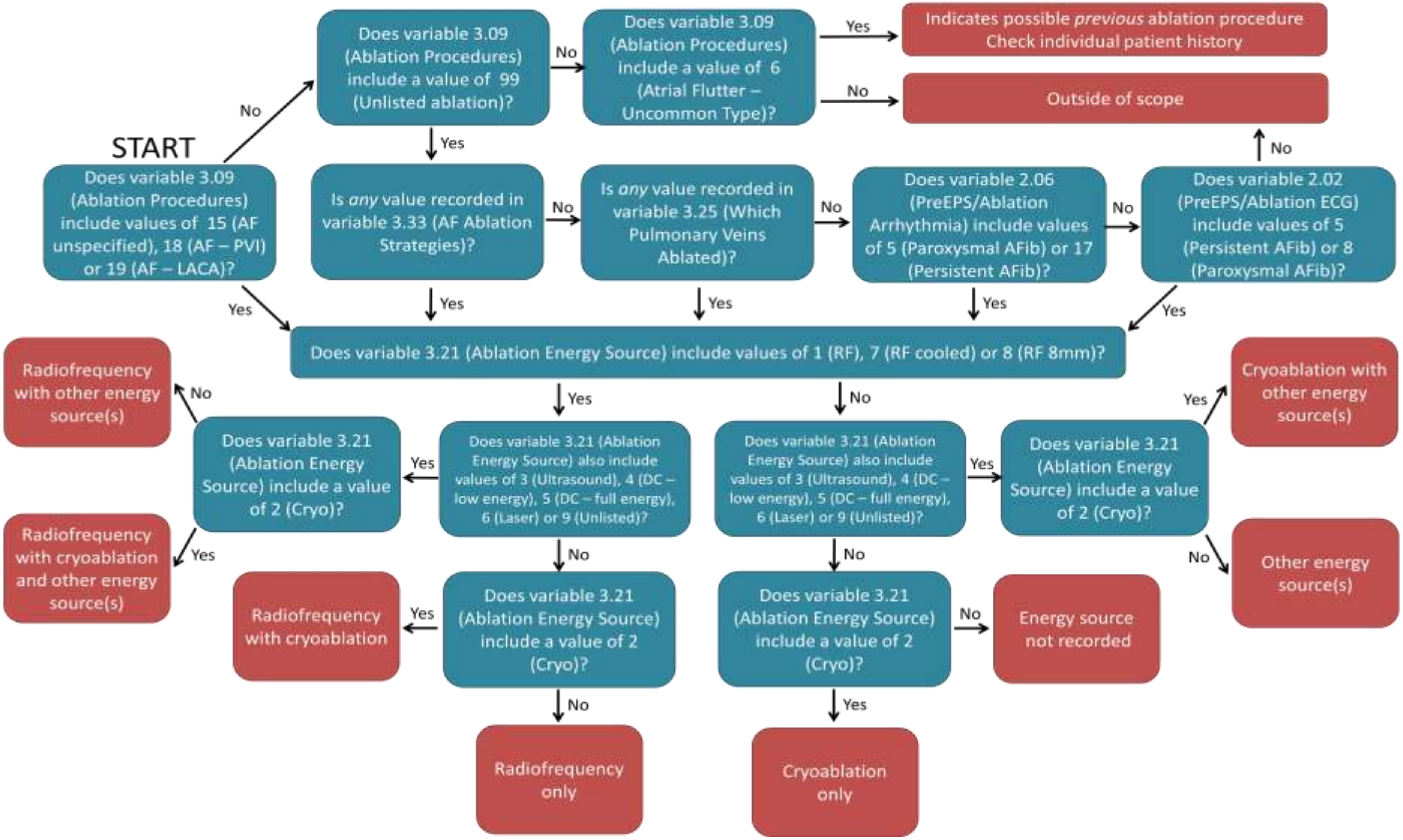

START
Does variable 3.09 (Ablation Procedures) include values of 15 (AF unspecified), 18 (AF – PVI) or 19 (AF – LACA)?
No
Does variable 3.09 (Ablation Procedures) include a value of 99 (Unlisted ablation)?
No
Does variable 3.09 (Ablation Procedures) include a value of 6 (Atrial Flutter – Uncommon Type)?
Yes
Indicates possible *previous* ablation procedure Check individual patient history
No
Outside of scope
Yes
Is *any* value recorded in variable 3.33 (AF Ablation Strategies)?
No
Is *any* value recorded in variable 3.25 (Which Pulmonary Veins Ablated)?
No
Does variable 2.06 (PreEPS/Ablation Arrhythmia) include values of 5 (Paroxysmal AFib) or 17 (Persistent AFib)?
No
Does variable 2.02 (PreEPS/Ablation ECG) include values of 5 (Persistent AFib) or 8 (Paroxysmal AFib)?
No
Yes
Yes
Yes
Yes
Yes
Does variable 3.21 (Ablation Energy Source) include values of 1 (RF), 7 (RF cooled) or 8 (RF 8mm)?
Yes
No
Radiofrequency with other energy source(s)
No
Does variable 3.21 (Ablation Energy Source) include a value of 2 (Cryo)?
Yes
Does variable 3.21 (Ablation Energy Source) also include values of 3 (Ultrasound), 4 (DC – low energy), 5 (DC – full energy), 6 (Laser) or 9 (Unlisted)?
Does variable 3.21 (Ablation Energy Source) also include values of 3 (Ultrasound), 4 (DC – low energy), 5 (DC – full energy), 6 (Laser) or 9 (Unlisted)?
Yes
Does variable 3.21 (Ablation Energy Source) include a value of 2 (Cryo)?
Yes
Cryoablation with other energy source(s)
Radiofrequency with cryoablation and other energy source(s)
Yes
No
No
No
Other energy source(s)
Radiofrequency with cryoablation
Yes
Does variable 3.21 (Ablation Energy Source) include a value of 2 (Cryo)?
Does variable 3.21 (Ablation Energy Source) include a value of 2 (Cryo)?
No
Energy source not recorded
No
Yes
Radiofrequency only
Cryoablation only

Figure 6 Draft algorithm for defining procedure type using data available within NICOR's register.

5.2.1 Comparators

A sample of individuals who underwent comparative treatment(s) will generally be identified using similar methods to those used to select individuals who underwent the intervention of interest.

However if a control group is needed, there are methods that can be used to choose an appropriate sample from a population, depending on the study design. These may be random samples, matched ‘exposed’ and ‘non-exposed’ groups or individuals, or a combination of various techniques.

5.3 Defining outcomes and covariates

As with the interventions, the ability to measure outcomes is dependent upon the nature and quality of the data recorded. Additional considerations here include the determination of end-points. For example, if a study were interested in discontinuation of medications, it may be challenging to recognise this through referring to routine data. Whilst a lack of prescribing data after a certain point in time may be evidence that a patient no longer requires the medication, there are many alternative scenarios that could equally explain this absence. The individual may have moved away, died, or be obtaining their medication ‘over the counter’ instead of by prescription. They might have previously built up a ‘stockpile’ of drugs, which they are now accessing rather than seeking new prescriptions. **If reliant on this type of outcome, then cautious interpretation is advised**.

5.3.1 Comorbidities

An advantage of using linked health data is that covariates such as age and gender are often available, and simple to incorporate in statistical analyses. A more challenging covariate is that of comorbidity, whereby one or more additional diseases or conditions exist alongside the primary condition of interest. In contrast with the carefully selected cases usually recruited to randomised controlled trials, ‘real life’ individuals can have multiple illnesses that might interact and modify the effects of the main intervention.

Various indices have been developed to summarise the extent to which an individual is affected by these comorbidities. The most well-known is the Charlson Index, which has been adapted for use with different types of data. Weightings are applied to a list of serious conditions, such as diabetes, dementia, heart failure and liver disease. The output is a numerical score, which is intended to estimate the mortality risk after ten years. This can be used to stratify study results or to adjust for confounding (a common source of bias).

Section K – Data preparation, analysis and reporting

1 Introduction

Once data have been released and linked, they need to be prepared for analysis. Much of the statistical analysis and reporting will be similar to that undertaken for other types of research projects.

The emphasis in this section will be on those aspects which may require a different approach because of the atypical characteristics of large, linked datasets and routinely-collected data. Unlike data from small clinical trials, it will not be possible to visually inspect the entire dataset to look for errors, inconsistencies or trends. Many of the processes undertaken at this stage will need to be automated – both those which identify issues, and those employed to modify the data. This will often require specialist programming knowledge and skills.

2 Preparing data for analysis

The preparation of data is likely to incorporate a number of different activities. **It is helpful for health informatics experts to work closely with those who will be running the statistical analyses and interpreting the results**. If those carrying out the data linkage and initial data preparation are not familiar with the research question and why the data need to be presented a certain way, confusion can arise and potentially misinterpretation of results. An example that we encountered with CALON was that sometimes a field (such as number of outpatient appointments) contained no data at all, when from an analytical perspective, it was more appropriate to record this as a zero in the dataset. A blank field would then imply missing data (data not available), whereas a zero would indicate that hospital records were available for that patient during that period, but that the patient did not attend.

Ideally, if the study has been well designed and data fields accurately defined in advance, then data preparation can be relatively straightforward. On the other hand, if data items in the received data file are not as expected, then resolution of issues may be complex and take many months. As previously mentioned in Section J, this is the main reason why we advocate viewing a sample of data in advance if possible, so that preparations can be made in advance.

2.1 Data cleaning

Issues with the quality of data may not be identifiable until the project data has been received. For example it may become evident that codes have not been entered consistently, or that multiple entries have been made within a field where only one value had been expected. These issues can sometimes be resolved, but may require some programming to create new fields, or to convert some codes from one configuration to another.

There will inevitably be some missing data. If this is a large proportion of the total possible population, then it may be appropriate to query how representative it is. Statistical techniques are available to manage missing data.

Once data have been irreversibly anonymised, it is not possible for researchers to go back to the original source to check details (as would be possible in clinical trial case report forms). The accuracy of the data has to be assumed. Unless it is a systematic error, then minor problems may not have a great impact on study results due to the sample sizes usually being very large in this type of work.

2.2 Data formatting

Data may have been received in a particular format that cannot be directly imported into the software that is intended to be used for analysis, requiring conversion. The particular arrangement of data within the dataset (or datasets) being used for analysis will depend upon the planned statistical analyses. It may be helpful to discuss this with a statistician prior to requesting data.

2.3 Application of exclusion criteria

Whether or not the data providers have applied the exclusion criteria prior to release of their data, it may be worthwhile checking that all ineligible patient records have been excluded. It should be relatively straightforward to write queries (computer programs used to retrieve information) to check this. For example, in CALON we noted that data for some patients under the age of 18 years had been provided, but the study was intended to be carried out on adults only. Identifying and excluding these records was a simple, but necessary exercise prior to conducting the analyses.

2.4 Calculations

New fields may need to be generated based on the application of algorithms, or calculated from existing fields. For example, it may be necessary to calculate time elapsed between two dates, or to account for censoring (such as where there are partial years of follow-up data). A more complex application would be the generation of Charlson Index values for comorbidity scores, which include various weightings (see Section J). See below for another example used in the CALON project.

3 Quality assurance

The quality of routine data in general was discussed in Section J. As well as examining the quality of data that have been received, it is necessary to check the quality of the processes used to prepare data, to provide assurance that the results are reliable and generalisable to the population of interest. These tests should be planned and carried out in a systematic way.

Those with expertise in the manipulation of large datasets are able to automate the quality checking processes, maintaining an audit trail of tests carried out and any alterations made to the datasets. As described below in the CALON project, using syntax for data processing facilitates quality checking. Similarly, well-documented and clear version control for datasets, syntax and outputs is of particular benefit if several people are accessing the same files, or for future reference. It might be helpful to make use of specialist version-control software, although this is not essential.

4 Statistical analysis

In general, the formal statistical analyses are likely to be conducted in much the same way as in other study types, and will differ depending on the study design. It is the characteristics of the data that really distinguishes this type of study from many others.

CALON

One of the challenges we faced with the CALON project was that it was designed to capture data for up to seven years for each patient – two years pre-procedure and five years of follow-up. Having decided upon a generalised linear mixed model for the statistical analysis, it was necessary for each of these years to be arranged as a separate row of data in the SPSS software. This meant that each patient had between one and seven rows of data. See table 3 to view part of the dataset structure for fictional patients.

Having multiple rows per patient added to the complexity of data preparation, particularly when using calculations to generate new fields. Due to the size of the dataset these processes needed to be automated. For example in determining survival, it was necessary to write syntax that would look for the latest record for each patient, taking the last date recorded from either primary or secondary care records. If data were only available for part of that final year, then the number of days survival in that year were added to the cumulative total for all other post-procedural records (but not from the pre-procedural rows). If any interim rows of data were missing, then 365 days were added to account for the fact that the patient must have survived for that 'missing' year in order to have later records. Each of these calculations required the computer to 'look' at the patient ID number in the rows above and below to confirm that the data belonged to the same patient. Sometimes this required the dataset to be turned upside-down!

We soon learned that writing syntax in full to document all data manipulation tasks (cleaning data, recoding, computing new variables and running analyses) was incredibly helpful for several reasons:

- It allowed us to maintain a thorough audit trail of what had been done to the data. Others were able to scrutinise the commands for quality assurance.
- Any time the SAIL team released a new version of the dataset, we were able to simply and quickly re-run syntax rather than manually clicking on drop-down menus and re-entering commands. In all, there were 12 releases of the CALON dataset.
- When tasks failed to execute correctly, we were able to examine the syntax to detect and correct the error.
- It was not necessary to maintain many copies of a dataset to record small changes. Only a few strategic versions needed to be saved.

Overall, writing and using syntax facilitated consistent and rapid reprocessing of data, and simplified quality checking activities.

Table 3 Example of data structure for two fictional patients. None of the information presented in this table is based on actual patient data.

CALON ID	Period	Procedure Date	Period Start Date	Period End Date	Further 191 data variables
123456	-2	12/11/2010	12/11/2008	11/11/2009	XXX
123456	-1	12/11/2010	12/11/2009	11/11/2010	XXX
123456	1	12/11/2010	12/11/2010	11/11/2011	XXX
123456	2	12/11/2010	12/11/2011	11/11/2012	XXX
123456	3	12/11/2010	12/11/2012	11/11/2013	XXX
789123	-1	06/05/2009	06/05/2008	05/05/2009	XXX
789123	1	06/05/2009	06/05/2009	05/05/2010	XXX
789123	2	06/05/2009	06/05/2010	05/05/2011	XXX

5 Reporting

Unlike tightly-controlled clinical trials and other studies in which data are prospectively collected under restricted conditions, routinely-collected data are likely to have been influenced by many unknown factors. It is therefore important to recognise and acknowledge the limitations of a study or its design, and to be cautious in interpretation of results.

When reporting the results of projects where explicit consent was not obtained, efforts must be made to limit the risk of statistical disclosure and preserve the confidentiality of individuals. This might be achieved by aggregating or suppressing results that apply to small numbers of individuals. Worth noting is that this could prevent thorough investigation or characterisation of rare events.

Another issue to be aware of when reporting the results of data linkage studies is that often the original data providers will have stipulated certain conditions of use, generally in a data sharing agreement (see Section H) or similar document. These requirements might mean that draft publications/presentations are made available to the data provider in advance, and/or that the source is appropriately acknowledged.

Depending on the focus of the study, researchers might disseminate study results within a particular clinical field, or may choose to share methodological advancements within a health informatics forum. Some such conferences are described in Section C. Similarly there are journals with themes relating to the use of healthcare informatics such as:

- BMC Medical Informatics and Decision Making
- International Journal of Healthcare Information Systems and Informatics
- Journal of Biomedical Informatics.

As previously noted in Section I, the RECORD guidelines are due to be published shortly. They are likely to be an important reference tool when reporting the results of studies using routinely-collected data.

Conclusions and key recommendations

1 Recommendations

Some of the key recommendations based on our experiences with CALON are as follows:

- Be aware of any developments in data protection guidance and legislation, and the implications they might have for research activities.
- Keep up-to-date with the latest advancements in data linkage methodologies through newsletters, conferences, events, networking, training opportunities and publications.
- Manage projects carefully to ensure that high priority activities are not neglected and to maintain progress in all concurrent tasks. It is helpful if one person takes a lead role and is involved with all organisations and oversees all project activities.
- Effective communication is key. Try to identify helpful contacts at each external organisation and develop relationships with them. Be persistent. Don't assume that individuals or committees within an external organisation will communicate well with one another. Keep a record of important communications, advice received and decisions made.
- Obtain support from those with expertise in health informatics and data linkage. Their insight into the complexities of using routine data, potential pitfalls and tried-and-tested solutions will be invaluable. Work closely with data analysts to ensure a mutual understanding of project objectives and how the data can be used to achieve those goals.
- The Farr Institute provides guidance medical research in the UK, whereas the Administrative Data Research Network is likely to be a key contact in facilitating social research.
- Engage a core steering group of researchers, analysts, health informatics experts, clinicians and patient representatives to guide the project.
- When designing a study, run a feasibility exercise first using an established data repository. Only attempt to create a new linkage if there is no alternative. Linkage might be necessary if different elements of the research question are only available within separate datasets. Note that incorporation of new data into an existing repository make take years of negotiation. An ad hoc temporary linkage has limited use and can be very costly.
- Once data sources that contain relevant information have been identified, contact the data provider at the earliest opportunity. Discuss the process that must be followed to obtain data, any specific requirements they have, and what the costs and likely timescales will be. Take advantage of any offer for collaborative working.
- Where possible, obtain sample data, or at the very least a data dictionary describing each field and the type of data it contains. Find out why the data were originally collected, who recorded them, how they were/are used, the date range available and the geographical coverage. Try to obtain reports on data quality and completeness, if available. If particular fields/variables are key to the success of a project, make sure they will be adequately defined by the available data.
- If planning to link data, make sure all datasets contain suitable patient identifiers. If setting up a new register, make sure that appropriate identifiers are included within the minimum (compulsory) dataset, to facilitate linkage to other datasets.
- If primary care data is of particular interest, consider making use of the SAIL Databank. The coverage of primary care data in Wales is much greater than that in England.

- When making applications for data, try to demonstrate prior responsible use of data by members of the project team (ideally through prior publications based on the same dataset). Collate documents relating to local data security measures to provide if requested. Read data sharing agreements carefully and negotiate details if necessary. Make sure informal applications are clearly marked as a draft.
- Information governance requirements are likely to differ depending on the study design and whether it is defined as research or not. Make use of the HRA decision tools when needed. Obtaining a favourable opinion from a research ethics committee may facilitate negotiations with some data providers. Other data providers may not require such approvals, for instance if the data are fully anonymised.
- Completion of 'safe researcher' training will provide assurance to data providers, and in some cases is a requirement prior to accessing data. Other training courses are available introducing data linkage methodologies and analysis of linked health data.
- Produce a protocol to summarise the study design, making sure that it adheres to any requirements that data providers have. Refine details in advance as far as possible, including listing operational definitions for each dataset field/variable. Keep data providers informed if changes to the study design are necessary at any stage.
- Examine code definitions and consider how they are intended to be used in a study. Watch out for potential unintentional misuse of codes or classifications.
- Be aware of the potential for errors and missing data in routine data, mitigating the impact where possible and acknowledging limitations if necessary.
- When reporting results, be cautious about interpretation and consider alternative explanations. Consult clinicians, patient representatives and analysts for different perspectives. Refer to the RECORD guidelines for minimum reporting requirements.

2 Conclusions

Of all of the above points, Cedar considers the most helpful lesson we have learned is to make use of established data linkage repositories (such as the SAIL Databank) wherever possible. Doing so as an initial feasibility exercise might answer the research question entirely, or help to refine it. If later proceeding with a new link to another dataset, the likelihood is that this initial feasibility stage will lead to better use being made of the linked dataset.

When first deciding whether data linkage is an appropriate method to use, it is important to consider whether the linked data is likely to be of sufficient benefit to justify the substantial amount of work that will be required to set it up. Is it possible to address the research question using an existing data resource instead?

The CALON project has shown that data linkage is a feasible methodology for answering research questions of benefit to the NICE Interventional Procedures programme, although we were unable to demonstrate this using data from England within the project timescale. The usefulness of linked routinely-collected data is evident from our report on the efficacy and safety of cardiac ablation procedures (Poole et al. 2014). There is also potential for use of linked routinely-collected data to provide evidence to support the work of other programmes at NICE.

References

Boyd P (2003) Health research and the Data Protection Act. Journal of Health Services Research and Policy 8(s1): 24-7.

Brooks CJ, Stephens JW, Price DE, Ford DV, Lyons RA, Prior SL, Bain SC (2009) Use of a patient linked data warehouse to facilitate diabetes trial recruitment from primary care. Primary Care Diabetes 3: 245-8 DOI: 10.1016/j.pcd.2009.06.004

Button LA, Roberts SE, Goldacre MJ, Akbari A, Rodgers SE, Williams JG (2010) Hospitalized prevalence and 5-year mortality for IBD: record linkage study. World Journal of Gastroenterology 16: 431-8 DOI: 10.3748/wjg.v16.i4.431

Dattani N, Hardelid P, Davey J, Gilbert R (2013) Accessing electronic administrative health data for research takes time. Archives of Disease in Childhood 98(5): 391-2 DOI: 10.1136/archdischild-2013-303730

European Commission (2012) Proposal for a regulation of the European Parliament and of the Council of the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf

European Commission (2014) Progress on EU data protection reform now irreversible following European Parliament vote http://europa.eu/rapid/press-release_MEMO-14-186_en.htm [accessed 27 May 2014]

European Commission (2015) Data protection day 2015: Concluding the data protection reform essential for the single digital market http://europa.eu/rapid/press-release_MEMO-15-3802_en.htm [accessed 6 February 2015]

Ferrante A, Boyd J (2012) A transparent and transportable methodology for evaluating Data Linkage software. Journal of Biomedical Informatics 45: 165-72 DOI: 10.1016/j.jbi.2011.10.006

Ford DV, Jones KH, Verplancke JP, Lyons RA, John G, Brown G, Brooks CJ, Thompson S, Bodger O, Couch T, Leake K (2009) The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Services Research 9: 157 DOI: 10.1186/1472-6963-9-157

Hall SE, Holman CDJ (2003) Inequalities in breast cancer reconstructive surgery according to social and locational status in Western Australia. European Journal of Surgical Oncology 29: 519-25 DOI: 10.1016/S0748-7983(03)00079-9

Health and Social Care Information Centre (2014) Code of practice on confidential information http://systems.hscic.gov.uk/infogov/codes/cop/code.pdf [accessed 12 February 2015]

James-Ellison M, Barnes P, Maddocks A, Wareham K, Drew P, Dickson W, Lyons RA, Hutchings H (2009) Social health outcomes following thermal injuries: a retrospective matched cohort study. Archives of Disease in Childhood 94: 663-7 DOI: 10/1136.adc.2008.143727

Jones KH, Ford DV, Jones PA, John A, Middleton RM, Lockhart-Jones H, Osborne LA, Noble JG (2013a) The physical and psychological impact of Multiple Sclerosis using the MSIS-29 via the UK MS register. PLoS One 8: e55422 DOI: 10.1371/journal.pone.0055422

Jones KH, Ford DV, Jones PA, John A, Middleton RM, Lockhart-Jones H, Peng J, Osborne LA, Noble JG (2013b) How people with Multiple Sclerosis rate their quality of life: an EQ-5D survey via the UK MS register. PLoS One 8: e65640 DOI: 10.1371/journal.pone.0065640

Jones KH, Ford DV, Jones C, Dsilva R, Thompson S, Brooks CJ, Heaven ML, Thayer DS, McNerney CL, Lyons RA (2014) A case study of the Secure Anonymous Information Linkage (SAIL) gateway: a privacy-protecting remote access system for health-related research and evaluation. Journal of Biomedical Informatics DOI: 10.1016/j.jbi.2014.01.003

Kelman CW, Kortt MA, Becker NG, Li Z, Mathews JD, Guest CS, Holman CDJ (2003) Deep vein thrombosis and air travel: record linkage study. BMJ 327: 1072 DOI: http://dx.doi.org/10.1136/bmj.327.7423.1072

Langan SM, Benchimol EI, Guttmann A, Moher D, Petersen I, Smeeth L, Sørensen HT, Stanley F, Von Elm E (2013) Setting the RECORD straight: developing a guideline for the REporting of studies Conducted using Observational Routinely collected Data. Clinical Epidemiology 5: 29-31 DOI: 10.2147/CLEP.S36885

Lewsey JD, Leyland AH, Murray GD, Boddy FA (2000) Using routine data to complement and enhance the results of randomised controlled trials. Health Technology Assessment 4: 1-55 DOI: 10.3310/hta4220

Lyons RA, Jones KH, John G, Brooks CJ, Verplancke JP, Ford DV, Brown G, Leake K (2009) The SAIL databank: linking multiple health and social care datasets. BMC Medical Informatics and Decision Making 9:3 DOI: 10.1186/1472-6947-9-3

Morgan KL, Rahman MA, Macey S, Atkinson MD, Hill RA, Khanom A, Paranjothy S, Husain MJ, Brophy ST (2014) Obesity in pregnancy: a retrospective prevalence-based study on health service utilisation and costs on the NHS. BMJ 4: e003983 DOI: 10.1136/bmjopen-2013-003983

National Information Board (2014) Personalised health and care 2020: Using data and technology to transform outcomes for patients and citizens – a framework for action https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/384650/NIB_Report.pdf [accessed 06 February 2015]

Patrick H, Sims A, Burn J, Bousfield D, Colechin E, Reay C, Alderson N, Goode S, Cunningham D, Campbell B (2012) Monitoring the use and outcomes of new devices and procedures: how does coding affect what Hospital Episodes Statistics contribute? Lessons from 12 emerging procedures 2006-2010. Journal of Public Health 35: 132-8 DOI: 10.1093/pubmed/fds056

Ploem MC, Essink-Bot ML, Stronks K (2013) Proposed EU data protection regulation is a threat to medical research: a suggested amendment would make most epidemiological and health research impossible. BMJ 346:f3534 DOI: 10.1136/bmj.f3534

Poole R, Dale M, Wilkes T, Rees A, Thayer D, Wang T, Ruschetti L, Carolan-Rees G (2014) CALON – Cardiac Ablation: Linking Outcomes for NICE – Efficacy and safety outcomes of cardiac ablation procedures. *Cedar External Assessment Centre - Internal report to NICE.*

Reilly R, Paranjothy S, Beer H, Brooks, Fielder H, Lyons R (2012) Birth outcomes following treatment for precancerous changes to the cervix: a population-based record linkage study. BJOG: An International Journal of Obstetrics & Gynaecology 119: 236-44 DOI: 10.1111/j.1471-0528.2011.03052.x

Roberts SE, Button LA, Hopkin JM, Goldacre MJ, Lyons RA, Rodgers SE, Akbari A, Lewis KE (2012) Influence of social deprivation and air pollutants on serious asthma. European Respiratory Journal 40: 785-8 DOI: 10.1183/09031936.00043311

van Staa TP, Dyson L, McGann G, Padmanabhan S, Belatri R, Goldacre B, Cassell J, Pirmohamed M, Torgerson D, Ronaldson S, Adamson J, Taweel A, Delaney B, Mahmood S, Baracaia S, Round T, Fox R, Hunter T, Guilliford M, Smeeth L (2014) Health Technology Assessment 18: 1-178 DOI: 10.3310/hta18430

White J, Carolan-Rees G (2013) Current state of medical device nomenclature and taxonomy systems in the UK: spotlight on GMDN and SNOMED CT. Journal of the Royal Society of Medicine Short Reports 4: 1-7 DOI: 10.1177/2042533313483719