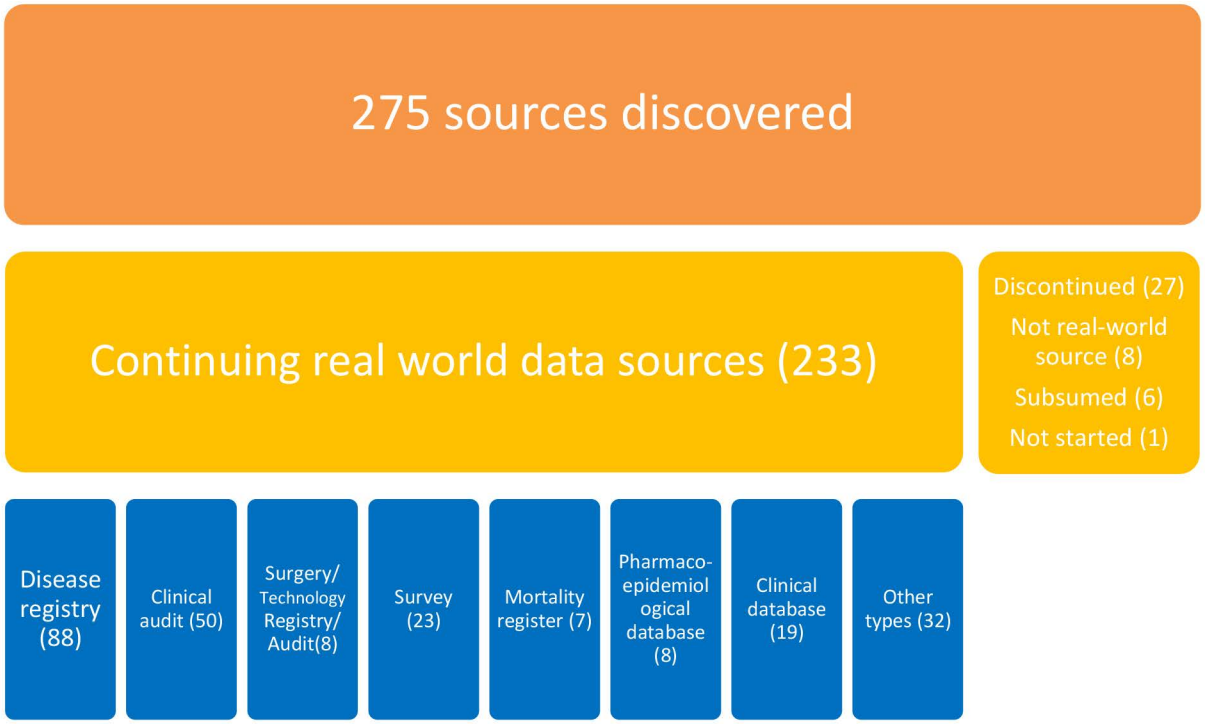
Institute of Education
EPPI-Centre

Identifying and appraising promising sources of UK clinical, health and social care data for use by NICE



Dylan Kneale, Meena Khatwa, James Thomas

EPPI-Centre
Social Science Research Unit
UCL Institute of Education
University College London

September 2016

The authors of this report are:

Kneale D, Khatwa M, Thomas J (EPPI-Centre, UCL Institute of Education).

This report was prepared by the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre), UCL Institute of Education for the NICE Science Policy and Research Programme. The views expressed here are those of the authors and not necessarily those of NICE, the EPPI-Centre, or any of the stakeholders involved in this research (see appendix for a list). ***All data were correct as of August 2015 and any errors and inaccuracies are the sole responsibility of the authors***. Responsibility for the views expressed remains solely with the authors.

Acknowledgements

We would like to acknowledge the contribution of all the expert stakeholders who participated in this project for their time and patience in inducting us into the complexities and possibilities of the real-world data landscape: Professor Nick Black (London School of Hygiene and Tropical Medicine), Professor Joanna Chataway (Open University/RAND Europe), Dr José-Luis Fernández (London School of Economics), Colin Flynn (Public Health England), Dr Shereen Hussein (Kings College London). Professor Martin Knapp (London School of Economics), Professor Jan Liliemark (SBU - Swedish Council on Health Technology Assessment), Dr Owen Nicholas (University College London), Dr Miriam O'Hare (Micron Group), Dr Louise Parmenter (Quintiles), Paul Ross (Social Care Institute of Excellence), Professor Tjeerd van Staa (University of Manchester), John Varlow (Health and Social Care Information Centre), Raphael Wittenberg (London School of Economics). We would like to acknowledge two real world data experts who also contributed and were based at a regulatory body and a public sector organisation.

We would like to acknowledge the thoughtful input provided by the internal NICE advisors for this project: Moni Choudhury (Science Policy and Research Analyst), Professor Sarah Garner (Associate Director of Science Policy and Research), Pall Jonsson (Senior Scientific Advisor), Jan Robinson (Science Policy and Research Programme Manager).

Conflicts of interest

There were no conflicts of interest in the writing of this report.

This report should be cited as: Kneale D, Khatwa M, Thomas J (2016), *Identifying and appraising promising sources of UK clinical, health and social care data for use by NICE*. London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London.

ISBN: 978-1-907345-93-7

Table of Contents

Executive Summary

Introduction

Understanding the topography of this real-world data landscape is of prime interest to NICE (National Institute for Health and Care Excellence) in this study, as well as gaining a snapshot of the key areas of debate in the field. Specifically NICE seeks to understand the way in which different real-world data sources could help to support NICE to realise its strategic objectives and to this end, NICE has identified five key uses of real-world data that can help the organisation meet its overall strategic objectives. These are to:

(a) **Research the effectiveness of interventions or practice** in real-world (UK) settings (e.g. through monitoring outcomes or proxy outcomes). Data could be used to inform the modelling of clinical and/or cost effectiveness outcomes as part of guidance production. Real-world data can also help to resolve uncertainties that have been identified in existing NICE guidance.
(b) **Audit the implementation of guidance.** For example, to assess the equity of implementation across different groups (including socioeconomic, geographic, demographic and groups differentiated by different diseases/health conditions); this may also form part of performance monitoring systems
(c) **Provide information on resource use** and evaluate the potential impact of guidance.
(d) **Provide epidemiological information.** For example prevalence/incidence of diseases, natural history, co-morbidities and information on current practice.
(e) **Provide information on current practice to inform the development of NICE quality standards**

EPPI-Centre Key Findings and Recommendations

Key Findings

- The real-world data landscape remains complex and heterogeneous and composed of sources with different purposes, structures and collection methods. This heterogeneity may increase with opportunities stemming from the incorporation of new technologies in data collection (current quality assured sources are limited in number)
- Some real-world data sources are purposefully either set-up or re-developed to enhance their data linkages and to examine the presence/absence/effectiveness of integrated patient care; however, such sources are in the minority. Furthermore, the small number that are designed to enable the monitoring of care across providers, or at least have the capability to do so at a national level, have been utilised infrequently for this purpose in the literature.
- Data that offer the capacity to monitor transitions between health and social care do not currently exist at a national level, despite the increasing recognition of the interdependency between these sectors.

- Among the data sources we included, it was clear that no one data source represented a panacea for NICE's real world data needs. This does highlight the merits and importance of data linkage projects and is suggestive of a need to triangulate evidence across different data, particularly in order to understand the feasibility and impact of guidance.

Key Overall Recommendation

- There exists no overall catalogue or repository of real-world data sources for health, public health and social care, and previous initiatives aimed at creating such a resource have not been maintained. As much as there is a need for enhanced usage of the data, there is also a need for taking stock, integration, standardisation, and quality assurance of different sources. This research highlights a pressing need for a systematic approach to creating an inventory of sources with detailed meta-data and the funding to maintain this resource. This would represent an essential first step to support future initiatives aimed at enhancing the use of real-world data.

Key Recommendations for NICE

Increased utilisation of existing sources beyond clinical databases:

- Making recommendations is difficult around the use of specific data sources. However, NICE's current use of real-world data differs substantially from the landscape with respect to its low utilisation of clinical audit, disease registry and survey data. Several of the datasets profiled in-depth highlight the potential of different sources of survey, clinical database and audit data.
- We also recommend that NICE further review its use of disease registry and audit data and engage in dialogue with collectors and depositors of these data to explore the utility of these types of data. Many sources of data available from disease registries and clinical audits are currently underutilised.

Investment in capacity and partnership building

- Use of real-world data requires substantial investment of resource that allows for the organisation to develop an in-depth understanding and experience of using different real-world sources. The extent of this undertaking should not be underestimated; any commitments and real-world data usage strategies should be matched by resources that allow for developing expertise in-house and in developing partnerships with data depositors and academic experts.
- Many of the data sources profiled either have active user groups or hold regular consultative exercises. NICE should further investigate these opportunities and capitalise on these.

Strategy and influence

- NICE has the potential to influence the availability of real-world data sources and good practice around the collection and utilisation of real-world data. This influence could be used to develop good practice around aspects such as obtaining informed consent from patients or obtaining investment around the

creation of data linkages. NICE should develop and publish an outward-facing policy around its use of real-world data which includes transparent means of influencing the state of the landscape, in order to ensure that sources continue to meet its organisational needs and to ensure alignment with national strategy. Exerting such influence could not only lead to benefits to NICE, but will have broader positive impacts across other stakeholders more widely, and could lead to improved patient and service user outcomes. This influence could also extend to developing quality standards around the way in which data are collected that can be shared across the sector.

- Care.data represents an initiative that could potentially meet many of NICE's real-world data needs. NICE should engage in discussions with the Health and Social Care Information Centre (HSCIC) to better understand and prepare for potentially using these data, while continuing to monitor whether and how the initiative overcomes challenges identified in earlier stages.

Understanding implementation

- Finally, while NICE is potentially able to monitor the implementation of guidelines using several sources, it may still lack information on the underlying mechanisms as to how or why guidelines succeed or fail in implementation. Starting its own programme of real-world data collection in the form of surveys of practitioners may be a way of understanding the mechanisms of un/successful implementation. Such an approach has been adopted elsewhere, for example by the Swedish Council on Healthcare Technology Assessment (SBU).

Green shoots

There are three key factors as to why the state of the real world data landscape should be regarded with some optimism for NICE and more generally.

1. Firstly, while data linkage and the capacity to research patient journeys is not at the point where many would desire, there are several examples where these efforts have been met with success and some of these have been met with a high degree of public acceptance. On a national level, the care.data initiative has restarted after a pause, and if these efforts succeed, they could meet many of NICE's real-world data requirements

2. Secondly, while we have been critical in the study about the representation of sources of patient reported outcomes, there are examples featured in the main report where patients have become more involved and have become gatekeepers to their own data (e.g. Salford Integrated Record), providing a possible model for the future. In addition, the ubiquity of smartphone technology and apps mean that ways of patients providing and managing their own information are increasing at pace.

3. Thirdly, methodological advances in the design and analysis of studies continue to ensure that real-world data becomes of greater utility for organisations, such as NICE, who wish to understand the implications of their decisions in real-world settings. These advances include the development of pragmatic trials using electronic health data which offer a balance between the methodological rigour of RCTs and the

generalisability of observational studies. Several UK based organisations and teams – some of which are represented among the expert stakeholders involved in the present study – are involved in driving these advances and it is likely that future studies will feature the results of these undertakings extensively in their findings.

What is real world data?

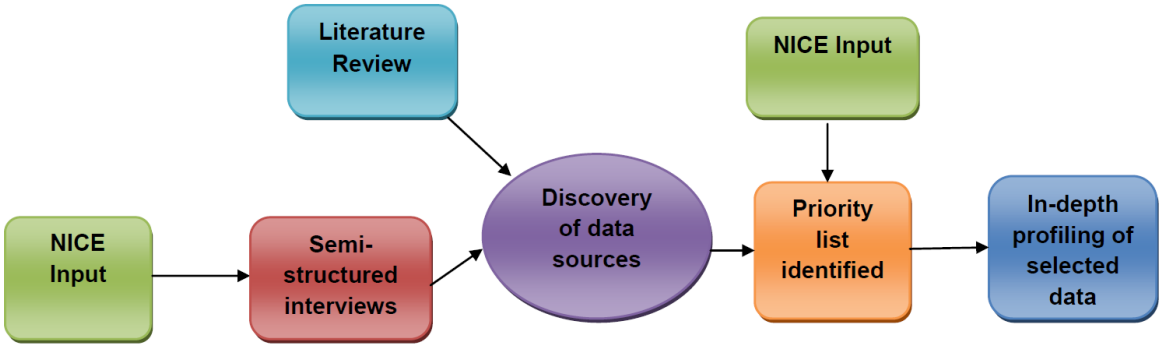
The definition of real-world data can be contentious and different stakeholders have different views as to what constitutes 'real-world' data. Real world data is defined in this report through two key tenets:

a. The collection of real world data reflects the usual care or treatment provided to populations of patients, service users or the public. This therefore excludes conventional Randomised Controlled Trials (RCT(s)) but could include other forms of RCT design, namely pragmatic RCTs[1].

b. Real world data provides enough depth to assess trends around everyday practice, service usage, or to assess the effectiveness of interventions and their outcomes.

To meet the needs of NICE, we do not pay close attention to sources of data that have limited geographic representation, and prioritise those sources with national or regional representativeness.

Study Approach



This study is focussed on identifying some of the available opportunities to NICE in terms of real-world data sources. To reflect the remit of NICE, in this report we consider data that spans clinical, public health and social care fields. To create a topographical map of real-world data sources we use:

- Data from interviews with expert stakeholders
- Data from studies discovered in the literature

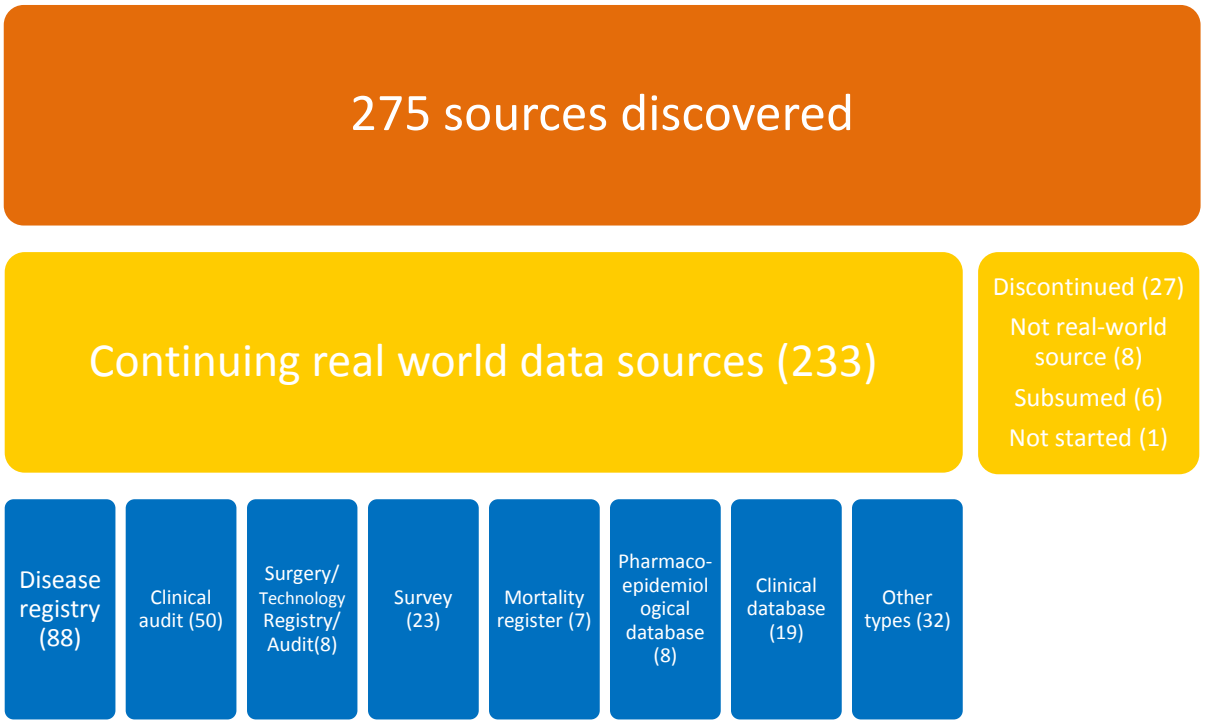
[1] Pragmatic randomised controlled trials aim to mimic real life conditions and test the effectiveness of a range of interventions that are known to be safe. They can be instrumental in understanding the relative effectiveness where there is no apparent clinical advantage/disadvantage among currently accepted treatment (1. van Staa T-P, Goldacre B, Gulliford M, et al. Pragmatic randomised trials using routine electronic health records: putting them to the test. Bmj 2012;**344**:e55.)

After drawing a long-list of sources, with the aid of NICE, we then identified those data sources that were not in current use, or were under-utilised, but were of interest to the organisation, and created an in-depth profile of eleven of these data sources.

Results - mapping of real-world datasets

In creating the map based on the literature and on interviewees' responses we discovered a total of 275 different sources of real-world data (figure 1), of which 233 are analysed further, being of most relevance to NICE. The remaining data sources were found to either have been discontinued (27) or subsumed into other studies (6 sources), were not actually real-world data sources (e.g. they were procedures or standards for application in the real world (8 sources)), or were at the protocol stage (one source).

Figure 1: Sources of real world data discovered[2]



How does NICE currently use real-world data?

Internally, NICE conducted a review of its use of real-world data across different teams. This review asked teams to name which data source was currently being used, how these data were accessed, processes employed for accessing data, associated costs; and a brief description of how the data were used:

The majority of data were found to be used, internally at least, to either (i) inform on the uptake of NICE guidance and/or explore use of medication (nine reports of usage could be described in this way (one represented future plans)); or (ii) for health economic modelling (twenty reports could be described in this way). Some of the data appear to

[2] It was also acknowledged that the HSCIC website held a great number of sources that could also be potentially profiled – future exercises could include a more detailed inventory of the HSCIC datasets.

support the development of quality standards particularly around safe levels of staffing (three reports could be described in this way) and there was one reported use of data for monitoring epidemiological and demographic trends. One dataset was described as being used to establish the effectiveness of interventions.

Evidence suggested NICE's internal use of real-world data differs from the real-world data landscape in the following ways:

- The use of clinical audit data by NICE does not match the widespread availability of these data
- The use of disease registry data by NICE does not match the widespread availability of these data
- Several sources of survey data are currently not being utilised internally
- Most of the data sources/functions currently in use appear to allow for cross-sectional analyses or repeated cross-sectional analyses only; patient-level longitudinal analyses appear to be conducted rarely
- Few datasets currently in use capture Patient Reported Outcomes
- NICE use an extensive array of different datasets to support understanding trends and the economic modelling of changes in prescribing trends
- There were no reports of NICE requesting additional data to be collected alongside standard data in any of the real-world sources
- Primary care data is based on The Health Improvement Network (THIN) data.
- Few of the existing sources allow for linkage across different services which a patient or service user may experience; however, these types of data were also underrepresented in the results of the mapping exercise
- For social care, none of the datasets described are directly sourced and explicitly focus on monitoring trends in private provision despite private provision being hugely important for the sector; these data were also underrepresented in the results of the mapping exercise

It should be noted that this review did not capture the multitude of data that are being used in work that NICE commissions from partner organisations.

What are some of the other broad debates and themes occurring in real-world data that NICE should be aware of?

There is no standard definition of real-world data and the term can be problematic

Often, real world data collected in primary care is viewed as a by-product of administrative or performance management activities. In social care, the definition of real world data tends to be broader from the outset and survey data was much more likely to be included in definitions. The breadth in the definition of real-world data was viewed as problematic by some and there is a need to clarify the distinction between 'real world data' and 'just all data'. Furthermore, the term 'real-world data' is not a familiar one across all disciplines.

Data for tracking patient and service user journeys are rare

Obtaining data that enables the tracking of patient and service user journeys and trajectories is fraught with difficulty. Much of this challenge is attributed to difficulties in being able to link data between sources using a common identifier, although considerable

efforts are underway to link across some datasets. Data that enable monitoring of transitions between health and social care are especially underrepresented.

Using real world data may involve using multiple datasets in addressing a single research question

Real world data's particular strength is the potential to provide the most complete picture available of the health and care status of the nation, and the services and interventions that are received in maintaining or improving health and care status[2]. These data are derived from a representative subset of the population and can provide a population-based snapshot of (i) illness or care needs; (ii) contacts with providers that take place; (iii) information on treatments or care packages, and (iv) ideally provide enough information on the outcomes of individuals. However, the real-world data landscape remains fragmented and the different elements that can provide a holistic understanding of patient and service user trajectories are stored in different sources with no means of linking.

Real-world data analysts will often encounter a trade-off between data that provides a depth of information on patient characteristics and data that provides a breadth of information on the services or interventions they receive. Real-world data projects often incorporate data from a number of different sources in order to overcome limitations within any given data source. In other real-world data projects, data from different sources are used in order to triangulate the findings and overcome potential concerns around representativeness or bias.

> *"There is a trade-off between having more information in terms of numbers and information in terms of breadth and depth of indicators. So survey data such as ELSA [English Longitudinal Study of Ageing] will give you a lot more in terms of quality of characteristics – income, wealth, needs, households' composition, service users etc. Certain outcomes will be much more limited on the other hand data from services; there will be thousands of cases in other sources - but much more limited - and the data and may not be of the same quality. We try to combine the data, look at patterns from both".*

Real world data complements the findings from randomised controlled trials

The main defining advantage of real-world data, besides apparent advantages in terms of cost, sample size and representativeness, is its (ostensibly) high external validity [1]. The external validity reflects both the delivery of an intervention to a group that is representative of the general population, but more crucially in the delivery of the control, which usually involves an alternative treatment regimen (best available alternative) as opposed to a placebo. While there is an expanding literature citing studies and study protocols that have been conducted using real-world data, interviewees (especially those from clinical backgrounds) emphasised that real-world data was not a replacement for/superseded the findings from RCT studies. Real-world data is prone to forms of

epidemiological bias unlikely to be replicated in findings from well designed and executed RCT studies [1 3]; however, as several interviewees pointed out, RCT study data can also be subject to bias, and some identified that observational data was subject to greater scrutiny despite its superior properties in terms of transparency, than RCT data are.

Future directions

Two themes emerged around future potential of real-world data. The first of these is around the expanding potential of pragmatic clinical trials (PCTs). Unlike traditional RCTs, PCTs are trials that take place within real-world environments and among representative samples of patients, thereby placing the focus on establishing the effectiveness of interventions, as opposed to their efficacy. Within a PCT, patients are randomised to receive an intervention or control treatment but the focus on mimicking real-world conditions means that, among other factors: (i) the control treatment provided often represents the best viable alternative already in place (as opposed to a placebo as can be the case in some RCTs), (ii) the patients randomised reflect the normal range of patients in terms of disease severity, comorbidity and demographic characteristics; and (iii) the measures of effectiveness collected as outcomes are valid and easily understood by a range of stakeholders, including clinicians, patients, policy-makers, and health commissioners. Real-world data collected through electronic health records was viewed as the basis for designing and undertaking a greater number of pragmatic trials (PCTs) and a number of real-world sources theoretically provide the means of implementing studies and monitoring outcomes in real-time. Evidence from PCTs is likely to be of substantial interest to NICE in establishing the effectiveness of interventions in real world settings while maintaining randomisation, thereby eliminating or at least substantially reducing the occurrence of channelling bias; the proliferation of real world data sources may facilitate this form of evidence to become increasingly frequent in the future.

A second theme that emerged was around new technologies stimulating new forms of real world data to be collected. Methods of collecting patient reported outcomes are shifting from paper to digital devices (smartphones and tablets): "*we have a lot of interest in technology where people get messages on their mobile phone to fill out symptoms, whether these are severe and so on. Uptake is very good and this type of model can be utilised for trials quite easily… where you have mobile phone technology sending information you don't have lots of paperwork… modern technology can help a lot with that. Also with ipads there is a strong movement to increases use in that.*"

Recommended specific sources to consider from interviews included:

1. Opportunities are available for assessing individual level service user outcomes through the Adult Social Care Survey (ASCS)
2. Understanding patient journeys and experiences using the broad scope of data contained within the National Cancer Data Repository
3. Assessing resource usage using National Minimum Data Set for Social Care (Skills for Care)
4. Exploring primary care practice using three of the large GP datasets
    - QResearch
    - Clinical Practice Research Datalink (CPRD)
    - The Health Improvement Network (THIN)
5. Understanding the contribution of risk factors to disease outcomes using the Whitehall II study
6. Exploiting the longevity and near-universality of the Myocardial Ischaemia National Audit Project (MINAP)
7. Examining epidemiological trends using Health Survey for England
8. Examining life course experiences on patterns of ageing using the National Child Development Study
9. Gaining a snapshot of social care and health service usage and needs of older people using the English Longitudinal Study of Ageing
10. Understanding epidemiological and care trends among households, including ethnic minorities, using Understanding Society
11. Monitoring Social and Health care trends, experiences and monitoring the implementation of standards using Care Quality Commission data and reports
12. Gaining an insight into patient experiences using Patients Like Me
13. Tracking data on patient journeys in integrated delivery networks: the potential of Scottish Health Informatics Programme (SHIP)
14. Exploiting Hospital Episodes Statistics Data as a multipurpose dataset
15. Exploring Epidemiological Trends using the Avon Longitudinal Study of Parents and Children (ALSPAC)
16. Using information from an integrated learning system through the Salford Integrated Record
17. Investigating the application of GP records
18. Understanding the effectiveness of interventions using National Joint Registry as a registry that is collecting longitudinal outcomes and patient reported outcomes
19. Understanding the effectiveness of interventions and monitoring the impact of guidance using the Sentinel Stroke National Audit Programme (SSNAP)
20. Understanding the effectiveness of interventions and monitoring the impact of guidance using the Renal Registry
21. Understanding the effectiveness of interventions and monitoring epidemiological trends using Adult critical care case mix programme (managed by ICNARC)
22. Harnessing the potential of cardiovascular audit and register data to address NICE's real world data needs
23. Data from the National Diabetes Audit; "the most advanced for long-term conditions"
24. Capturing genetic information on biomarkers in the UK Biobank
25. Calculating cost effectiveness based on data from the Personal and Social Services Research Unit
26. Mental Health and Learning Disabilities Data Set
27. Understanding trends in screening rates, healthcare and epidemiology using Quality and Outcomes Framework (QoF) data
28. Understanding epidemiological trends and measuring the effectiveness of interventions using the UK Inflammatory Bowel Disease Audit
29. Harnessing the potential of audit data to address NICE's real world data needs through clinical audits conducted by the Royal College of Surgeons Clinical Effectiveness Unit
30. Data used to populate NICE's Return on Investment Tools
31. Data from private health providers and insurers

Further sources that were shortlisted for consideration based on the literature/input from NICE were: (i) Care.data (ii) Prescribing observatory for Mental Health. It was also acknowledged that the HSCIC website held a great number of sources that could also be potentially profiled – future exercises could include a more detailed inventory of the HSCIC datasets.

Specific data profiled

From the earlier long-list of 30+ datasets, a selection of eleven data sources was chosen for in-depth profiling based on input from the NICE steering group. Datasets were prioritised if they were not in current use by NICE and where they appeared to meet some of the broader gaps in usage or addressed any of the themes emerging from the interviews. A template was developed to capture the properties of different sources according to their suitability for NICE's intended usage.

Focus on the potential utility of different datasets for NICE

All the profiled data sources are likely to have some utility to NICE dependent on the research question and making a specific recommendation around use is challenging as this is very much dependent on the context and the focus of the research question. The following section summarises the utility of the different sources for NICE. A full description of each dataset is provided in the main report.

English Longitudinal Study of Ageing (ELSA)

- ELSA has been used to establish the effectiveness of interventions at a population level using observational methods, for example in a cost-benefit analysis of cataract surgery among ELSA respondents [4]. ELSA may be less suitable for establishing the effectiveness of more specialist interventions/practice, or establishing how interventions/practice vary among minority groups.
- ELSA can be used to determine the implementation of guidance through examining broad population-level temporal changes in the receipt of common interventions or practice. For example ELSA data were used to examine shortfalls in care for chronic conditions using set quality indicators [5]. Without further linkages, ELSA data may be less suitable for explaining the underlying mechanisms around the implementation of guidance, beyond patient/service-user characteristics.
- ELSA data can be used to provide information on some aspects of resource use, for example how many people receive common interventions and can be used to establish how access may vary by individual patient characteristics.
- ELSA data can be used to establish self-reported levels and determinants of many age related conditions and non-communicable diseases and more broadly information on lifestyle behaviours and attitudes among older people.
- ELSA data may be less suitable for establishing the incidence/prevalence/outcomes of very uncommon diseases/conditions/interventions.

Community Mental Health Survey (CMHS)

- The CMHS data have been used to monitor the implementation of guidance, for example in monitoring the implementation of guidance aiming to strengthen support for service users during times of turnover in staffing [6]. The data have also been used to draw together guidance around expected standards of care [7]. There may also be potential to use the data to monitor different aspects of resource usage.
- The focus of the survey is on service user experiences and there is less information on outcomes following receipt of different forms of care, limiting the utility of the data with respect to establishing the effectiveness of interventions. The data are less suitable as a tool for monitoring epidemiological patterns in mental health.

Clinical Practice Research Datalink (CPRD)

- CPRD data have utility for NICE through the flexibility in being able to collect additional fields. CPRD data are also available to medical researchers based outside UK universities potentially expanding the pool of potential partners with which NICE could work in using the dataset. The long established nature of CPRD (based on General Practice Research Database) means that several retrospective studies could also be potentially conducted using these data.
- There are numerous examples where CPRD (and GPRD) data have been used in studies that cover all of NICE's intended uses of real-world data. For example, CPRD have been used to evaluate changes in cancer diagnostic intervals following the introduction of NICE guidance [8]. Given the potential to draw large samples, studies could be implemented that examine the epidemiology/outcomes/implementation of rare or less common conditions and procedures. Unlike survey-based sources, for example ELSA and HSE, and in the absence of further data collection, there is potential to examine only a limited range of patient-level intrinsic factors, although these may be sufficient for many studies.
- Data linkages will expand the utility of CPRD data for NICE; current linkages include those with MINAP data, National Cancer Intelligence Network data and HES data. Area level data are also available including Index of Multiple Deprivation data and Townsend deprivation scores [9]. Further data linkages are planned.

QResearch

- QResearch is of interest to NICE for many of the real world data uses identified by NICE, but access appears to be restricted to research consortiums led by academic institutions. Nevertheless, given the substantial potential of these data, NICE could consider ways of developing research projects based on QResearch data led by universities.
- There is potential for QResearch data to be used in studies that cover all of NICE's intended uses of real-world data. The use of QResearch data in developing risk prediction scores may also be of interest to NICE, potentially around forecasting and modelling future disease burden.
- Given the potential to draw large samples, studies can be implemented that examine the epidemiology/outcomes/implementation of rare or less common conditions and procedures. One example is a study of peanut allergy, where a prevalence rate of 0.51 per 1000 patients in the UK was estimated [10].
- The study depositors state that QResearch data are suitable for case control studies designed to examine risk factors for onset of disease, cross sectional surveys, cohort studies and sample size calculations (for non-observational studies) [11].
- As is the case for all three large primary care databases, there is potential to examine only a limited range of patient-level background characteristics, although these may be sufficient for many studies.

The Health Improvement Network (THIN)

- There are numerous examples where THIN data have been used in studies that cover all of NICE's intended uses of real-world data. For example, THIN data have been used to examine equity in access to cancer screening among people with Intellectual Disabilities compared to those without across different types of cancer [12].

- Data linkages expand the utility of THIN, and THIN data have been linked with Hospital Episodes Statistics (HES) data, providing potential for studying continuity in care between primary and secondary care. A number of patient postcode-based socioeconomic, ethnicity and environmental indicators are available to researchers including Townsend deprivation quintile scores.
- Overall there is a wide scope for analysing data reflecting outcomes and experiences of morbidity and mortality at primary care level, as well as trends in the care and treatment provided. These data can also be linked to HES data allowing for potential tracking of patient journeys between primary and secondary care. As is the case for all three large primary care databases, there is potential to examine only a limited range of patient-level intrinsic factors, although these may be sufficient for many studies.
- THIN data have utility for NICE through the flexibility in being able to collect additional fields and the potential to conduct research based on free-text fields. THIN data are also available to medical researchers based outside UK universities potentially expanding the pool of potential partners with which NICE could work with in utilising real world data.

National Minimum Data Set for Social Care (NMDS-SC)

- NMDS-SC is a specialist dataset suitable for monitoring trends in the social care workforce. This data can potentially help NICE to understand workforce capabilities and undertake preliminary work to understand the feasibility of implementing new standards and guidance in social care settings.
- The data may be suitable to examine changes following the implementation of NICE guidance at a workforce level in terms of indicators such as pay, training or necessary skills. They may also be useful in helping to set benchmarks and develop quality standards around workforce capacity and skills. The data have also been incorporated into calculations of resource use in the literature [13]. While the data do not provide insight into epidemiological trends per se, they do provide insight into the workforce preparedness for responding to epidemiological challenges, such as dementia [14].
- As social care outcomes are not collected in NMDS-SC, it is unlikely that these data are suitable for researching the effectiveness of interventions and practice.

Health Survey for England (HSE)

- HSE was suggested in the context of monitoring epidemiological trends, although the potential usage extends beyond this purpose alone and potentially HSE data can be used to gain an understanding of trends over time in terms of resource utilisation, trends in social care needs and usage, trends in lifestyles and social determinants of health, and some trends in prescribing, service usage and attitudes to health. With regards to researching the effectiveness of interventions, in the absence of data linkages, there may be more limited potential to measure the effectiveness of interventions or changes in practice. Examples where data have been linked to explore later outcomes include an examination of fruit and vegetable intake and mortality [15].
- The survey data may be of great utility for NICE in gathering contextual information critical in the assessing feasibility of different forms of guidance aimed at public health and social care challenges. The data also have the added advantage of being relatively easy to obtain for further secondary data analysis and are free to use.

- There is scope for auditing the implementation of guidance through examining change in practice at a population level; one of the strengths of HSE data in doing so is the ability to examine social or medical inequalities in the implementation of guidance. Some HSE information may be suitable in providing information for the development of NICE quality standards and these data may be particularly useful where the standard is based on meeting a certain level of patient satisfaction or experience.

Adult Social Care Survey

- ASCS is a survey of users' satisfaction with the care that they receive. Such data can be used in forming guidance that is based on user experience and patient reported outcomes. There may be limited scope for undertaking secondary analysis of the individual service user data without further permissions being sought. Nevertheless, the detailed reports and tables produced may allow for gaining a good level of understanding of aspects of service user satisfaction with their care and broader aspects of wellbeing.
- With regards to measuring the effectiveness of practice, while it may be possible to undertake repeated cross-sectional studies and examine the impact of changing practice on user experiences, fully assessing the effectiveness of interventions through measuring longitudinal changes at a service-user level will be challenging with these data. However, it may be possible to assess whether guidance is being implemented, particularly around service user satisfaction or service user reported experiences, through analysing change (for example at a Local Authority (LA) level).
- With regards to using the data as an epidemiological tool, the study provides a snapshot of general health trends and social care needs but among a population who are receiving LA assistance for these health needs (the sample design represents a caveat around the applicability of the data). There may be scope for the data to be used to form quality standards around social care experiences and trajectories – for example around information advice and guidance received by older people in accessing care.

Salford Integrated Record (SIR)

- SIR was suggested as a source of data that may have the potential to overcome the limitations of other data source and examine patients' integrated care pathways. The potential of the data for research purposes is likely to be in the process of being realised and there are comparatively few publications using these data in the literature; the data may have been used initially to mainly facilitate clinical decision-making and performance management. Perhaps one of the most appealing characteristics of the data, given the current climate around the use and ethics of electronic health records in medical research, is the high degree of patient involvement and the ability of patients to access their own records.
- The data hold substantial potential for improving patient care. The integration of primary and secondary care data allows for research tracking patient outcomes across care providers (through examining Integrated Care Pathways (ICP)). One initiative using the data in this way is the Collaborative Online Care Pathway Investigation Tool that is being used to examine missed opportunities in patient care – that is where primary prevention opportunities were missed which could lead to adverse health outcomes. This initiative is focussed on modelling the circumstances and frequency of variance between idealised ICP and the actual care provided 16.

Prescribing Observatory for Mental Health (POMH-UK)

- One of the key criteria for choosing a topic focus of the POMH-UK is that the topics are relevant for monitoring the implementation of NICE guidelines. This has direct relevance to one of the intended uses of real-world data by NICE. An example of study directly assessing the implementation of NICE guidance can be found in a study of renal and thyroid functioning among patients who are prescribed lithium 17.
- The utility of the data for other more research-focused or evaluative activities, for example in assessing the effectiveness of interventions or monitoring epidemiological trends, may be more limited. The data are not widely used in the literature and it is unclear the extent to which these data are made available for re-analysis, reflecting their primary function as a quality improvement tool. Nevertheless, there are several important questions that could be addressed for NICE as there may be potential to understand whether practice/outputs have changed over time. In addition, this source represents one of the few specialist sources of real-world data on mental health encountered.

Care.data

- If successfully implemented, care.data would make a substantial contribution to the real-world data needs of NICE and other organisations. The data could allow for establishing the long-term effectiveness of interventions through the capacity to track patient journeys through primary and into secondary care as standard, something that rarely occurs as standard in real-world data projects and sources. Uniquely, it could also potentially, allow for insight into patterns of social care and their relationship clinical and public health data.
- At the time of writing it is too early to tell the extent to which care.data has been able to overcome the challenges encountered, particularly around consent and conditions around data usage. The results of the pathfinder exercise will offer further insight into the viability of the whole project; the majority of testing in pathfinder areas is due to begin later this year.

Section 1: Introduction, Background and Methods

This report is split into five main sections:

- The first section introduces NICE's ambitions for the use of real-world data and some of the properties of real-world data that can enhance its robustness, as described in the literature. This section also includes details about the methods used to collect information in this report.
- The second section presents the results of the map of real-world data sources and compares the current use of real-world data by NICE with the landscape of available sources.
- The third section introduces additional information from interviews with expert stakeholders who provide further commentary on the state of the real-world data landscape including fundamental issues around what should be considered real-world data.
- A fourth section includes data profiles from ten selected datasets that are not currently being used, or are only being used in a limited way, by NICE.
- The fifth section provides a short summary from (eleven) selected real-world data sources that have been profiled in-depth as well as including another organisation's experiences of using real-world data.
- The remainder of the report comprises appendices including a full list of the 275 'data sources' that were initially discovered and summarily appraised.

1.1 Introduction

A decade ago interest in real world health and clinical data peaked with the publication of several systematic enquiries aimed at mapping the breadth and depth of sources of real world data [2 3 18 19]. These reviews each had a different foci, and spanned a few years apart, but were consistent in highlighting that sources of real world data were plentiful and fragmented, having different strengths, weaknesses and idiosyncrasies; and were of varying utility to decision-makers at different stages of decision-making and monitoring processes. Previous reviews highlighted that many sources were long established: Newton and Garner's review of disease registries included a description of Norway's Leprosy register, set-up in 1856 and which continued until 1973, and held 8000 records including four from patients still alive when discontinued. Arguably other forms of real world data, such as the record of cholera cases developed by John Snow in 1849, have an equally long-standing history. One of the first documented real-world health data projects might be found in the work of John Graunt, who analysed 2500 bills of mortality in London in 1663 mortality and cause of deaths over 50 years in an effort to understand patterns of bubonic plague [20]. Over 350 years later, new sources of real world data continue to be established, for example through the establishment of the UK's first biobank [21] , and efforts to link and consolidate existing sources of real world data continue. Alongside these advances we've also seen greater investment in streamlining access points for real world data, for example through the establishment of the Health and Social Care Information Centre and its continued work. However, while there are some who laud the UK's commitment and usage

of real world data in health policy [22 23], there are others who question the pace of advances [24]. For a number of decades there have been various calls for greater systematic collection and utilisation of data collected routinely in healthcare. For example, in 1980 the establishment of the Körner committee, and the recommendations made and implemented, ensured that greater impetus was placed on the collection of data to improve patient care. More recent national efforts to improve, expand and link health and care data have attracted a number of critiques, reflecting concerns around ethical issues, consent and confidentiality [25 26]. The flagship care.data project was recently given a red flag of 'low confidence' by the government's own watchdog and has been beset by concerns around confidentiality, privacy and data ownership [27]. The widespread concerns around this particular project have meant that public confidence in the use of real-world data has been eroded, possibly leading to suboptimal clinical advancements as well as high levels of distress around the use of personal data [28].

The extent to which concerns around the security and ethics of real-world data collection have served to limit the proliferation of real-world data sources, against a backdrop of expanding opportunities around the capture and analysis of real-world data, are relatively unknown. In addition, the type of real-world data that has attracted much recent controversy – linking of primary care data and its use for commercial and analytical purposes – is but one model of real-world data collection and real-world data use. As we reveal in this report, the landscape around real-world data is complex, and while different sources may have similar, if not identical, data collection designs, they may nevertheless have very different policies and be at different stages when it comes to data linkages and data usage. New insights and perspectives have also prioritised different aspects of patient care, and there is now a growing focus on collecting information seldom included in reporting systems in the past, for example patient reported outcomes [29], as well as genomic data. Therefore the extent to which the conclusions of previous reviews of data sources still stand, in that UK "healthcare systems, despite requiring information on whether they provide the right interventions well and fairly to the right people, tend to have poor information systems" [19; p65], are subject to debate and review. More recent reviews suggest that issues reflecting the lack of unity in data collection systems and fragmentation of sources identified in previous studies persist, hindering the potential contribution of real-world data [30], although new sources continue to be established that offer potential insights into areas of health and care where comparatively little has previously been known[31].

Furthermore, health care commissioning and delivery structures have changed dramatically in recent years, necessitating a broader scope including public health and social care sources of real world data in the current study, as well as clinical sources. The commissioning of health care has shifted to fall within the control of GPs as part of clinical commissioning groups. Public health surveillance and services in England now fall under the remit of Directors of Public Health based in Local Authorities, although with bridge links to health care. The planning, commissioning and delivery of social care continues to be based in Local Authorities, although the need for integration is recognised across the spectrum, and this is being reflected in new healthcare structures being developed, for example the recent shift to Local Authority partnership control of health care services in Manchester [32]. NICE's own remit since 2012 now includes all three areas (social care, public health, clinical healthcare). Notwithstanding the changes in demography,

epidemiology and care needs, attitudes and behaviours, and advances in health and social care technologies; these changes in health and social care delivery structures will also drive changes in data collection patterns and needs. The shift of services to more localised structures may offer opportunities for the development of networks of best practice in the collection of real-world data. Political support for greater utilisation of real-world data for research and improving patient outcomes was common across all the prospective candidate parties at the 2015 general election [33], indicative of substantial political will to strengthen and utilise real-world data.

Unlike previous reviews in this arena, this current study takes a less systematic approach and adopts a broad brushed approach, not being confined to any one given particular type of real world data, and encompassing health, social care and public health real world data. To help meet the challenge of addressing the breadth, this study takes an expert-driven approach, using semi-structured interviews with a number of experts in the field as a source of data, alongside creating map of the literature, the production of a case study, and supplementary exploration of a select number of these sources as case studies. Additionally, unlike previous reviews, the current study is being undertaken with the specific data needs of NICE driving the focus. The remainder of this chapter is focussed on (i) introducing the methods used to produce this report; (ii) introducing principles of good-practice and strengths of real-world data; and (iii) introducing principles of good practice and strengths of real-world data that reflect NICE's specific real-world data needs.

Defining real-world data

The definition of real-world data can be contentious and different stakeholders have different views as to what constitutes 'real-world' data. Real world data is defined in this report through two key tenets:

a. The collection of real world data reflects the usual care or treatment provided to populations of patients, service users or the public. This therefore excludes conventional Randomised Controlled Trials (RCT(s)) but could include other forms of RCT design, namely pragmatic RCTs[3].

b. Real world data provides enough depth to assess trends around everyday practice, service usage, or to assess outcomes.

[3] Pragmatic randomised controlled trials aim to mimic real life conditions and test the effectiveness of a range of interventions that are known to be safe. They can be instrumental in understanding the relative effectiveness where there is no apparent clinical advantage/disadvantage among currently accepted treatment – in the case of a pragmatic RCT, instead of any one of these treatments being administered arbitrarily, the treatments are prescribed through random allocation (1. van Staa T-P, Goldacre B, Gulliford M, et al. Pragmatic randomised trials using routine electronic health records: putting them to the test. Bmj 2012;**344**:e55.)

To meet the needs of NICE, we do not pay close attention to sources of data that have limited geographic representation, and prioritise those sources with national or regional representativeness.

This definition therefore allows for the inclusion of clinical databases, disease and case registries, workforce registries, administrative records, HTA registers, surgical registers, surveys, clinical audits, population registries, service records and censuses to be included as sources of real world data (see glossary).

1.2 Methods

Research questions

In order for NICE to fully capitalise on the increasing availability of data, there is a need to better understand the extent and relative strengths and limitations of current sources and how they may change in the future. Therefore in this report we address two main research questions:

a. **What are the main sources of clinical, health and social care data that are currently available that match the needs of NICE?**
b. **What are the main features, strengths and limitations of some of these available health, clinical, social care and public health datasets for NICE?**

Two supplementary research questions include:

c. **What are the main debates around using routinely collected and other real-world health, social care and public health data (as identified by experts and other sources)?**
d. **What can NICE learn from another agency's experience of using real-world data?**

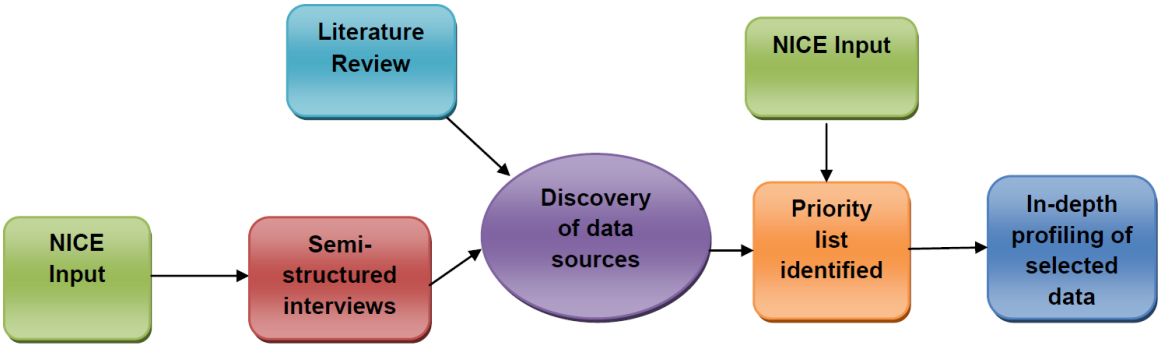
With respect to the first question, those sources that are not in current use by NICE are those which are of greatest interest.

Overall approach

In this project we attempt to map and appraise the availability of clinical, health, social care and public health data sources using an expert-led approach and supplemented through a review of the literature and website searching. We discover and map the most relevant available sources of clinical, health, social care and public health data for NICE, mainly relying on evidence gathered from expert interviews to shortlist a number of these sources to profile in greater depth. This forms an expert-driven map of the most relevant sources of data that fit within the needs of NICE that is supplemented by explorations of the literature. This means that the results presented here (i) do not form and do not attempt to form a comprehensive map of different sources of real-world data; (ii) are weighted towards the real-world data needs of NICE, and therefore other health and social care stakeholders may prioritise these sources differently; and (iii) are focussed on those sources not in current use by NICE. While the research and views here are independent of

NICE, we received input through suggestions for expert stakeholders to interview and we worked collaboratively with a NICE steering group to prioritise those datasets for in-depth profiling.

Figure 2: Process showing how different data sources were selected for in-depth profiling



Semi-structured interviews

The purpose of the interviews is to both ensure we are made aware of different sources of routine data as well as considerations in assessing its quality for use by NICE in its work. In the interviews we explore:

- Knowledge of different current UK sources of clinical, health and social care data most relevant for use by NICE in its work (which alongside desktop research will help us to ensure that we are able to map out the main ***relevant*** sources)
- Knowledge of strengths and limitations of the named sources (in the context of the needs of NICE).
- Current and emerging debates around the collection, storage, analysis and utilisation of routine and non-routine health, clinical and social care data
- Knowledge of examples of organisations with a similar remit to NICE that use real-world data in a way which could be considered as best practice

We were also interested in the way in which stakeholders perceived the overall quality and breadth of the landscape, their experiences of using different datasets, and the strategies that they used to identify the strengths and limitations of these data, and in some cases their future plans for using the data. Interview participants were initially contacted either by email or telephone and meeting times set up. The list was drawn up partly on the basis of exploring the existing literature and partly through input from NICE. Interviews were semi-structured in format, and conducted using a topic guide (which was modified according to area of expertise of the expert stakeholder (see Appendix 3 for an example). The topic guide was developed through summarily exploring the literature. The interview schedules were structured around considering the usefulness of data for NICE

and the properties and characteristics needed to assess the effectiveness of health technologies and practice/interventions; the quality and accessibility of data; debates around the availability of data, and the relative merits of real-world data for NICE. While a focus of the interviews was on establishing the main sources of data that could be used by NICE, we also probed for less well-known sources. Interviews lasted between 40-90 minutes, although most were around 50 minutes, and were partially transcribed by a member of the research team. We did include verbatim quotes where appropriate to support some themes. Responses were analysed thematically, moving from attributing codes to the data to exploring themes, although maintaining a focus on mapping the real world data landscape.

Literature search and case study

We reviewed the available academic and grey literature using a structured search and then complemented by forward and backward citation searching. We initially conducted a search of PubMed using search terms reflecting geography, scope (health and social care) as well as terms reflecting real-world data (see box below).

Box 1: Search string used to search for literature

(((((("uk"[Title/Abstract]) OR "england"[Title/Abstract]) OR "wales"[Title/Abstract]) OR "scotland"[Title/Abstract]) OR "northern ireland"[Title/Abstract]) AND and AND ((health) OR "social care") AND and AND (((((("real world data") OR "clinical database") OR "social care data") OR "health data") OR "clinical audit") OR "disease registry") OR "routine data" OR "administrative data") in PubMed

This search uncovered 743 results (700 after the removal of duplicates). To build on previous reviews that took place 2004-2006 [2 18 19 34] and to minimise the number of discontinued databases that may be included in the results, a further filter was put on the search to exclude literature published before 2005; this reduced the number of results to 548. All these references were imported into EPPI-Reviewer 4 and screened on title and abstract. In the case of 384 of these studies, the information in the title and abstract was sufficient to decide on whether the study included real-world data (or not) and the name of the real-world data source included. In 164 of the included studies, further information was extracted from the full text.

From the interviews, we sought to identify a case study of an organisation that used real-world data with a particular emphasis on identifying 'good practice'. We sought to identify organisations that had used data to (i) directly assess technologies/interventions; (ii) address research gaps emerging from guidance; (iii) examine the extent of implementation of guidance; (iv) as part of economic analyses of research gaps or implementation; (v) measure regular performance and monitoring activities (e.g. medicine metrics). To produce the case study, we undertook: (a) a semi-structured interview with a representative of the organisation; (b) collection of/request for archival/documentary evidence of the (positive) impact of the use of real-world data e.g. in terms of studies undertaken by the organisation to address research gaps stemming from guidance or in the production of guidance; (c) collection of/request for evidence of the broader impact using

real-world data in the organisation's activities, as well as (d) evidence on the criteria used to measure strengths or weaknesses of the data and strategies used to limit the impact of any weaknesses. Not all of these elements were populated in our case study (section 5).

1.3 What makes for good real-world data (for NICE)?

The following criteria/principles have been identified from the literature and the interviews conducted as being important; some are idealised and rarely feature in the available data but are nevertheless important considerations. Few data sources are likely to fulfil all desired properties listed. Commonly, studies will triangulate data from different sources in addressing research questions, such as Vinogradova and colleagues' [35] study, which examined data from two primary care datasets (CPRD and QResearch); this can be one method of minimising some of the complications of using real-world data sources. Using multiple sources offers wider coverage and richer information in combination, which can maximise the capture of information on the effectiveness of an interventional procedure [36-38]. However, the capacity to do this is very much dependent on the research question/purpose of the real-world data project.

Furthermore, even the most methodologically robust real-world data sources will have a complementary, as opposed to competing, role alongside other forms of evidence. This includes evidence from conventional experimental studies (i.e. conventional RCTs) as opposed to observational or pragmatic trial studies; the former remaining essential to establish efficacy and safety of interventional practice and the latter remaining the only source of evidence of effectiveness.

Principles/Criteria for selecting/assessing sources of real-world data:

Clearly defined aims: Data sources should have clearly defined aims that NICE can subscribe to; this appears particularly important in the present time as the purpose and use of real-world data is in debate.

High levels of validity; clear case definition: Validity of constructs in real-world data connotes the degree to which indicators measure the conditions they purport to measure. This is usually achieved through implementation of standard measures/frameworks for measures; in registry data the problem can be framed differently in terms of clear case definition (and inclusion in the data), although the principles remain the same. The use of standard coding frames, for example ICD-10[4] (International Classification of Disease) as used in clinical databases can be one way in which validity of constructs/case definitions can be improved; standard measures are also implemented in social care data (e.g. ASCOT; ADLs; IADLs) although there may be greater variety and subjectivity in some of these measures.

Clear case definition is also dependent on the diagnostic tests for the condition – e.g. even a common condition like asthma actually has no definitive diagnostic test but rather a set of symptoms and reactions to medication that provide diagnosis [39].

[4] http://apps.who.int/classifications/icd10/browse/2015/en

Some constructs may have good validity but through being less stable, can have lower levels of reliability. For example constructs such as body mass index (BMI) have lower levels of reliability in real-world data than more stable constructs (e.g. genetic markers) as they are subject to rapid change. This can be problematic for some combination of uses and sources of real-world data; for example if real-world data from primary care was used to select overweight/obese patients for a cohort study based on height and weight measures from the previous appointment, some of the patients included may no longer be classed as obese/overweight.

Representative of populations and setting: A theoretical strength of real-world data is its high generalisability in comparison with data from RCT studies. This can only be achieved if data are representative of populations and settings [18]. This can include being sectorally representative (e.g. representative of public and private providers)[34]. As a further condition, analytically and to enhance generalisability, real-world data should also be representative of 'large' populations or a large fraction of cases if appropriate [3]. For example, localised registers of rare diseases may be less useful than national data for NICE both from an analytical (sample size) perspective and in terms of generalisability. For registry data in particular, different geographically-based registries may have been established independently originally, but may have developed a unified framework so that instead of representing geographic areas, they may have switched focus to a particular clinical or epidemiological specialism. Where fragmented data sources exist, the ability to combine data with other local datasets can reduce the possibility of type II statistical errors in particular from occurring. Data linkage in general is often problematic with real-world health data both analytically and in terms of ethical considerations.

Representativeness in terms of disease stages: This indicates the degree to which real-world data include patients with different stages of a disease or condition, for example people with a spectrum of needs from mild to severe. In some cases, a register or database may only include patients with certain stages of a disease or condition, but in others a database/register may purport to include all cases but may in fact systematically underrepresent people at a certain stage or with a certain severity of a disease or condition for example, [40].

Clear ethical frameworks aiming for informed consent as gold standard: Real-world data sources should adhere to recognised codes of ethics and as a gold standard should include an opt-out clause or obtain explicit consent [18]. In practice this can be difficult, and while there are some who speculate that obtaining consent results in the introduction of bias; others view the lack of transparency around consent as a limiting factor on the extent of real-world data sources and their scope. Researchers find that most UK patients would allow their data to be used by the NHS for research purposes but do have concerns about the selling of information for commercial gain as well as in the potential misuse of data [41].

Dynamic and adaptable (multipurpose): Dynamism in the goals and ambitions of real-world data collectors can lead to innovation and investment in the data. For example a new database on suicide in Wales – the Suicide Information Database-Cymru is being developed by researchers based on routinely collected health and social care data linkage

[42]. This is the first to allow the tracking of healthcare pathways and contact with different providers in the UK of suicide cases.

Flexibility to add new data and ideally to add new data on a request/trial basis: This is particularly important as the choice of indicators can be driven solely by the indicators available [43], as opposed to those addressing a research question of clinical importance. Consider the situation where pain is an outcome; real-world sources may not capture changes or improvement without additional Quality of Life indicators and Patient Reported Outcomes being collected.

Flexibility to implement different study types: For example the flexibility to implement the different commonly occurring epidemiological study types: prospective cohort, retrospective cohort, case control and nested case control studies etc. Different study types using same data can find different results – e.g. link between statins and common cancers found in one nested case-ctrl study [44] but not in a retrospective cohort study [45] of GP data (QResearch and THIN data respectively); therefore flexibility to implement different designs and triangulate the results is important in real world data.

Capable of being linked to other data (or potential) for tracking patient outcomes: A common identifier allowing linkage of different dataset is one of the priority goals/movements towards improving the potential of real-world data. Data linkage in this way allows for the tracking of patient journeys through health systems (or potential for) and is a key ingredient in forming population health systems [46 47]. Some work in this area is underway in linking different clinical databases (HES and CPRD) and clinical databases and clinical registers.

Collection of data reflective of real-world conditions (scope): One of the clear advantages of real-world data is the 'real-worldness' of it and being able to create a robust economic and budget impact argument is an important part of this [37]. Real-world data presents an opportunity to understand the effectiveness of interventions and to collect information that provides unique insight into establishing the effectiveness of interventions, including a sufficient breadth of intrinsic variables and a sufficient breadth of prognostic variables.

This includes considering whether the scope is broad enough to consider events beyond morbidity and mortality.

Sensitivity: Due to their breadth and longitudinal nature, some sources of real-world data are better suited than others to enable analysts to distinguish between pre-existing comorbidities from complications of treatment or iatrogenic diseases [36].

Transparent sampling frames/recruitment processes: This is an important consideration in order for researchers to understand and report upon potential selection effects in the data. Ideally, databases and registers will seek complete coverage of the cases within scope although this may not always be achieved [18 48], and can lead to selection effects (bias) where there are systematic differences between cases included in datasets and those who are not. Different forms of real-world data may be vulnerable differentially to selection effects. Some sources suggest registers are more exposed to selection bias than clinical databases [40 48], but participation in some of the major primary care databases is

also voluntary at practice and patient level [2], which can compromise representativeness. For NICE, the priority should be in the use of real-world data where the impact of possible selection effects can be ascertained – e.g. through clear study documentation – and potentially adjusted for where necessary.

Where sampling frames have been used (e.g. in survey data), the construction of probabilistic weights and/or response weights and/or clear identification of strata and sampling units may be necessary to make inferences to wider populations.

Uniformity in data collection procedures: While this may not be possible for some sources of real-world data, use of standard tools for measurement can be one way to attempt to ensure the validity of measures [1]. Uniformity with other sources will aid comparability, including with international data [43]. Within a dataset, guides and training around data collection and input may provide information needed to assess the uniformity of data collection processes.

Steps taken/can be taken to minimise common form of bias: Due to the mainly non-randomised nature, forms of real-world data may be accompanied by a degree of bias in the collection or design of the data (as well as in the analysis). While researchers may not be able to fully resolve these forms of bias, an ability to minimise or understand extent of common forms of bias specific to real-world data is important. This is particularly the case for those types of bias that do not ordinarily occur in RCT designs; see [49]; these forms of bias include:

a. ***Confounding by indication, channelling bias*** - where the underlying risk profile differs for those receiving a treatment compared to those who are not [50]. This is not insurmountable, or at least the impact can be minimised in part, e.g. use of instrumental variables, propensity score matching, risk stratification and other methods can be used to minimise impact on estimates – but all these techniques depend on having broad scope of data collection and the techniques themselves are not infallible and can introduce a degree of subjectivity [51]
b. ***Protopathic bias*** – establishing the actual sequence between diagnosis and treatment; establishing the sequencing and dosage of different treatments a patient may be receiving
c. ***Additional forms of allocation bias***
d. ***Recall and forms of reporting bias*** see [36] – indicates situations where the data collection itself compromises the information collected for a number of possible reasons including respondents' inabilities to remember or give accurate information, acquiescence, tendencies to give extreme or rounded values, as well as a tendency towards social desirability
e. ***Tolerance bias (pharmacoepidemiology)*** see [50] – indicates whether there is scope for the detection of medicines for comorbidities (plus over the counter medicines)
f. ***Information/detection/observer/ascertainment/assessment bias*** – indicates the possibility that those with vested interest in doing so may underreport; also possibility that misclassification occurs see [50]
g. ***Selection bias*** – see selection effects above

Responsiveness: Whether the data are responsive enough to record changes in treatment or care package including the timing and sequencing of switching treatments.

Quality Assured: Robust sources of real world data will produce detailed quality assurance reports and quality assurance measures. These steps can adhere to common standards such as the National Statistics Quality Mark [34].

A degree of data user involvement: Steering groups or user groups are one way in which real-world data can maintain a balance between what is pragmatically feasible and what data users require for research purposes [34]; these can help users to influence the direction of real-world data sources.

Accessible and ad hoc analyses feasible: Data should be accessible and with little unnecessary administrative burden. The ability to access granular data (e.g. patient/practitioner level data) is a clear advantage for re-analysis and addressing new research questions. Access should be balanced between the need to undertake clear, hypothesis driven research and the ability to undertake exploratory data analysis to explore trends, e.g. around dose response. Accessibility also includes access to clear data documentation and a low, or at least purposive, administrative burden for access – i.e. for example the process of gaining ethical approval is beneficial for researcher and data holder. Accessible data is also data that is low cost or free to access for research purposes.

Support and training: The most useful sources of data may be those where support is provided in the form of manuals, helpdesks, training. The existence of a data usage library or archive in order to understand who is already using the data and why can also help to avoid duplication in studies.

High data quality: This is an implicit aim of all real world data sources; high quality data can include data that include few logical inconsistencies and the ability to identify duplicate cases. [2 47]

Secure and confidential: Secure and confidential data protection procedures are essential in ensuring that the data are ethical; for NICE they can also minimise the reputational risks and internal risks of using real-world data

Good levels of stability: The better the stability/longevity of the resource, the more expansive the research questions that can be addressed. In addition, use of a data source can represent an investment in itself in terms of access and training and using a data source with low levels of stability or likely longevity can represent a risk in itself.

Granularity of treatment/disease data: ICD disease codes and treatment codes may not ordinarily capture the specificity/granularity required for some conditions; although such a depth of detail may be more likely to be found in registry/audit data [36] (this consideration is highly dependent on the research question in mind and for most types standard disease codes may suffice)

Timely: Timeliness is a theoretical advantage of real-world data and a criteria in assessing real-world data should be its temporal relevance [23]

1.4 What properties do NICE need of real-world data?

The ambitions of NICE in using real-world data are summarised in five key functions. Each of these functions can entail different data requirements and are discussed below.

1. Research the effectiveness of interventions or practice in real-world (UK) settings

Real-world data would enable NICE, having first established the efficacy of interventions through experimental evidence, to assess the transferability of conventional RCT findings in real-world practice. An example of an intervention where its effectiveness has not been established in real-world settings is presented in Box 1, taken from NICE guidance around managing headaches among young people [52].

Box 2: Example of need for researching the effectiveness of interventions in real-world settings

CG150/1: Is amitriptyline a clinically and cost effective prophylactic treatment for recurrent migraine? Why this is important:- Effective prevention has the potential to make a major impact on the burden of disability caused by recurrent migraine. There are few pharmacological agents that have been proven to prevent recurrent migraine. Amitriptyline is widely used, off-label, to treat chronic painful disorders, including migraine. Inadequate evidence was found in the review for this guideline for the effectiveness of amitriptyline in the prophylaxis of migraine. A double-blind randomised controlled trial (RCT) is needed to assess the clinical and cost effectiveness of amitriptyline compared with placebo. The definition of migraine used should be that in the International classification of headache disorders II or this guideline. Outcomes should include change in patient-reported headache days, responder rate and incidence of serious adverse events. If amitriptyline is shown to be effective, it will widen the range of therapeutic options, in particular for people in whom recommended medications are ineffective or not tolerated.

While the recommendation states that a double-blind RCT is needed to assess the cost-effectiveness of the drug amitriptyline, such a design would not be feasible using real-world data, and in particular, such a design would not give a full account of the cost-effectiveness of amitriptyline in the real-world, since patients suffering from recurrent migraine would unlikely be offered a placebo, but instead the best available alternative treatment. Investigating the effectiveness an intervention in real-world settings would usually require:

i. Clear determination of the target population (who received the intervention) as well as a method of identifying controls (in order to establish comparative effectiveness)
ii. Reliable measurement of the desired outcomes (and therefore usually requiring longitudinal design)

iii. Satisfactory range of intrinsic measures included (i.e. to control for unwanted sources of variability, to assess the stability of intervention effects across groups and to enhance generalisability)
iv. Satisfactory range of additional prognostic variables (i.e. to control for potential confounding around treatment options)
v. Given that effectiveness can refer both to clinical effectiveness as well as effectiveness in terms of resource use or cost, other measures around inputs and outputs including staffing or patient may be required

A 'gold-standard' study design for measuring the effectiveness of interventions in real-world settings might involve conducting a pragmatic RCT [1], although in practice this may not always be possible and the identification of a retrospective cohort may be a more common study design for assessing the effectiveness of interventions. Where there is a pre-defined outcome of interest, particularly a rarer outcome, a case-control study may be more appropriate. In rarer situations, a case-series design may suffice (where no control group is identified). Regardless of the study design, the properties identified above are those that are required of real-world data in order to measure the effectiveness of interventions.

2. Audit the implementation of guidance

Real-world data would enable NICE to assess the degree to which guidance is implemented and how this implementation may vary across different groups; for example including socioeconomic, geographic, demographic and groups differentiated by different diseases/health conditions. An example of guidance is provided below where the drug Naftidrofuryl oxalate was identified as the recommended option for the treatment of intermittent claudication in people with peripheral arterial disease, and other drugs (Cilostazol, pentoxifylline and inositol nicotinate) were not recommended for this purpose.

Box 3: Example of need for real-world data for auditing the implementation of guidance

TA2323: Cilostazol, naftidrofuryl oxalate, pentoxifylline and inositol nicotinate for the treatment of intermittent claudication in people with peripheral arterial disease

1.1 Naftidrofuryl oxalate is recommended as an option for the treatment of intermittent

claudication in people with peripheral arterial disease for whom vasodilator therapy is considered appropriate after taking into account other treatment options. Treatment with naftidrofuryl oxalate should be started with the least costly licensed preparation.

1.2 Cilostazol, pentoxifylline and inositol nicotinate are not recommended for the treatment of intermittent claudication in people with peripheral arterial disease.

1.3 People currently receiving cilostazol, pentoxifylline and inositol nicotinate should have the option to continue treatment until they and their clinicians consider it appropriate to stop.

NICE identified that this guidance was being broadly implemented in a review conducted in 2014, and that prescriptions of naftidrofuryl oxalate had increased since 2011 and the evidence suggested that prescriptions for cilostazol, pentoxifylline and inositol nicotinate had decreased using the ePACT (Electronic Prescribing Analysis and Cost Tool) and Hospital Pharmacy Audit Index information [53]. However, the data also showed that reductions in cilostazol, pentoxifylline and inositol nicotinate prescriptions had tailed off following an initial decline in the first quarter of 2012. ePACT data allow for the potential for examining some practice level characteristics although, should NICE require, alternative sources of real-world data may also shed light on the characteristics of prescribers and patients who are not using naftidrofuryl oxalate. To gain summary insight as to whether guidance is being implemented, at minimum the data should allow for:

i. Clear determination of the intervention(s) (which may include implementing a diagnostic tool as well as 'treatment')
ii. Means of establishing change over time that with relatively low levels of recall or other forms of bias (e.g. a repeated cross-sectional design for data capture)
iii. Measuring outcomes is not a pre-requisite although the ability to capture patient/prescriber/institutional level characteristics and outcomes will enable a deeper understanding. As most NICE guidance provides criteria for use rather than a binary recommendation, then fuller scrutiny of patient records is likely necessary for full auditing of the implementation of guidance [54].

Auditing the nuances of some recommendations may be particularly difficult using routine data. For example, Box 3 below shows that in the case of overweight and obese adults, it is recommended that GPs raise issues of weight loss in a respectful and non-judgemental way. Auditing this element of the recommendation would require subjective data from patients and clinicians which may not be captured in the majority of existing real-world data sources (although provisions could be made). Researchers who have audited the implementation of NICE guidelines in the past have used a number of different taxonomies of real world data including data from existing clinical audits, clinical databases, disease registries but also commonly have designed specific data collection tools (survey based methods that are often also included within audits) as well as undertaking interviews and examining locally sourced case series data [54-56]. Furthermore, there is also an important distinction between observing trends and attributing changes to the implementation of NICE guidelines. In other words, while a real-world data project auditing the implementation of guidance may find a trend in practice, attributing this to the impact of the guidance itself and not to other factors (such as other developments in professional wisdom) is difficult.

Box 4: Example of need for real-world data for auditing the implementation of guidance

PH53/6 Recommendation 6: Refer overweight and obese adults to a lifestyle weight management programme

GP practices and other health or social care professionals who give advice about, or refer people to, lifestyle weight management programmes (see Who should take action?) should:

- Raise the issue of weight loss in a respectful and non-judgemental way. Recognise that this may have been raised on numerous occasions and respect someone's choice not to discuss it further on this occasion.
- Identify people eligible for referral to lifestyle weight management services by measuring their body mass index (BMI). Also measure waist circumference for those with a BMI less than 35 kg/m^2. Consider any other locally agreed risk factors.
- For funded referrals, note that:
    - programmes may particularly benefit adults who are obese (that is, with a BMI over 30 kg/m^2, or lower for those from black and minority ethnic groups) or with other risk factors (comorbidities such as type 2 diabetes)
    - where there is capacity, access for adults who are overweight should not be restricted (that is, for people with a BMI between 25 to 30 kg/m^2, or lower for those from black and minority ethnic groups) or with other risk factors (comorbidities such as type 2 diabetes)
    - there should be no upper BMI or upper age limit for referral.
- Provide information on programmes available locally, where possible, taking people's preferences and previous experiences into account. Be clear that no programme holds the 'magic bullet' or can guarantee long-term success.
- Refer people to a group rather than an individual programme if they express no preference because, on average, group programmes tend to be more cost effective.
- Ensure people who are overweight or obese who are not referred (for whatever reason) have an opportunity to discuss and reconsider attending a programme in the future. Discuss making a follow-up appointment at an agreed date (for example, in 3 to 6 months). Provide them with sources of information about how to make gradual, long-term changes to their dietary habits and physical activity levels (for example, NHS Choices).
- Give people the opportunity for a re-referral, as necessary, because weight management is a long-term process. Use clinical judgement, taking into account the person's circumstances, previous experiences of weight management and commitment to change.

3. Provide information on resource use and evaluate the potential impact of guidance in changing resource use

A summary examination of health or social care resource usage may include information on the type of health or social care resource utilised (e.g. a hospital or residential home stay, a GP visit or drug prescribed) as well as the amount of resource used. To gain summary insight for resource use at minimum the data should allow for:

i. Clear determination of the intervention(s) delivered and the resources employed (which may include implementing a diagnostic tool as well as 'treatment').
ii. Means of establishing the amount of resource used.
iii. Resource data for developing/informing guidance may need to include the data necessary to calculate Quality Adjusted Life Years (years lived in 'perfect' health).
iv. Measuring outcomes is not a pre-requisite although the ability to capture epidemiological information, health outcomes (including potentially patient reported outcomes) and cost information may be essential for a fuller cost-effectiveness study (as opposed to studies investigating trends in resource use).

4. **Provide information on epidemiological trends**

Box 5: Example of need for real-world data for understanding epidemiological trends

CG53/3 What is the prevalence and incidence of CFS/ME in different populations? What is the natural course of the illness?
Why this is important: Reliable information on the prevalence and incidence of this condition is needed to plan services. This will require well-constructed epidemiological studies across different populations to collect longitudinal data to predict outcome, and to calculate the economic impact of loss of work or education. We recommend that these questions are answered using a mixture of:

- cross-sectional population studies, including people with different levels of disease severity from all ethnic groups and social classes
- longitudinal cohorts of people with CFS/ME, and population cohorts to assess the incidence and prognosis of CFS/ME in a previously normal cohort

Epidemiological trends may refer to summary trends (usually over time) in prevalence or incidence of diseases or conditions; more in-depth studies may also explore associations in terms of socio-demographic or socioeconomic inequalities, regional inequalities, or may seek to explore the impact of different exposures and their associations with the likelihood of developing diseases or conditions compared with staying healthy. While epidemiology usually refers to ill-health, the same principles may extend to social care if we consider the risks of developing different levels of care needs, for example. In order to monitor trends alone, outcome data may not necessarily be a pre-requisite, although as the research recommendation in box 5 outlines, there is usually a desire both to study the incidence/prevalence of disease as well as prognostic data.

At minimum the data should allow for:

i. Clear determination of the disease/condition/state through diagnostic tests OR standardised disease coding OR a battery of symptomology data collected with which to determine a (secondary level) diagnosis. This may therefore require information from a variety of sources, for example blood assay or radiography scan data.
ii. Satisfactory range of intrinsic measures included (i.e. to assess associations across groups and to control for unwanted variability)

iii. Satisfactory range of exposure/risk variables (i.e. to assess the impact of different behaviours and exposures in elevating or reducing risk)
iv. Satisfactory range of additional prognostic variables (i.e. to assess disease course)
vi. Where relevant, reliable measurement of the desired outcomes (and therefore requiring longitudinal design)

5. Provide information on current practice to inform the development of NICE quality standards

Quality standards produced by NICE 'describe high-priority areas for quality improvement in a defined care or service area. Each standard consists of a prioritised set of specific, concise and measurable statements' and an example is provided in Box 6 (see [57]). The use of real-world data for this purpose does not entail monitoring the adherence to quality standards, but instead involves investigating where different thresholds of practice are associated with measureable outcomes, be these in objective clinical or social care outcomes or in terms of patient reported outcomes. The data requirements for this purpose are likely to mirror those for the first purpose in requiring:

i. Clear determination of the target population (who received the intervention) as well as a method of identifying controls (in order to establish comparative effectiveness)
ii. Clear determination of the timing, sequencing and level of inputs provided as part of the intervention
iii. Reliable measurement of the desired outcomes (and therefore usually requiring longitudinal design)
iv. Satisfactory range of intrinsic measures included (i.e. to control for unwanted sources of variability, to assess the stability of intervention effects across groups and to enhance generalisability)
v. Satisfactory range of additional prognostic variables (i.e. to control for potential confounding around treatment options and determine other care being provided)

Box 6: Example of quality standard

NICE quality standard [QS81]: Quality statement 1

People with suspected inflammatory bowel disease have a specialist assessment within 4 weeks of referral.

1.5 Summary of data requirements based on the intended use of real-world data by NICE

Table 1: data requirements based on NICE's intended use of real-world data

	Essential for basic study	Desirable for more in-depth study	Potential study designs for basic study
Research the effectiveness of interventions or practice in real-world (UK) settings	i. Clear determination of the target population; method of identifying controls i. Reliable measurement of the desired outcomes i. Satisfactory range of intrinsic measures included v. Satisfactory range of additional prognostic variables	i. Given that effectiveness can refer both to clinical effectiveness as well as effectiveness in terms of resource use or cost, other measures around inputs and outputs including staffing or patient may be required	Longitudinal data ideal e.g. cohort (prospective or retrospective cohort) Repeated cross-sectional data may provide indicative evidence of hospital/ unit/centre trends
Audit the implementation of guidance	i. Clear determination of the intervention(s) i. Means of establishing change over time	i. Ability to capture patient/prescriber/institutional level characteristics	Repeated cross-sectional data may suffice
Provide information on resource use and evaluate the potential impact of guidance in changing resource use	i. Clear determination of the intervention(s) delivered and the resources employed i. Means of establishing the amount of resource utilised	i. Resource data for developing/informing guidance may need to include the necessary data to calculate Quality Adjusted Life Years (years lived in 'perfect' health) ii. The ability to capture epidemiological information, health outcomes (including potentially patient reported outcomes) and cost information may be essential for a fuller cost-effectiveness study	Repeated cross-sectional data may suffice
Provide information on epidemiological trends	i. Clear determination of the disease/condition through diagnostic tests OR standardised disease coding OR a battery of symptomology data collected with which to determine a (secondary level) diagnosis. i. Satisfactory range of intrinsic measures included i. Satisfactory range of exposure/risk variables	i. Where relevant, reliable measurement of the desired outcomes (and therefore requiring longitudinal design)	Longitudinal data ideal e.g. cohort (prospective or retrospective cohort) or case-control studies

	v.	Satisfactory range of additional prognostic variables	
			Cross-sectional/repeated cross-sectional data may provide indicative evidence of overall trends
Provide information on current practice to inform the development of NICE quality standards	i.	Clear determination of the target population (who received the intervention) as well as a method of identifying controls (in order to establish comparative effectiveness)	Longitudinal data ideal e.g. cohort (prospective or retrospective cohort)
	i.	Clear determination of the timing, sequencing and level of inputs provided as part of the intervention	
	i.	Reliable measurement of the desired outcomes	Repeated cross-sectional data may provide indicative evidence of hospital/ unit/centre trends
	v.	Satisfactory range of intrinsic measures included	
	v.	Satisfactory range of additional prognostic variables	

Section 2: (Findings I) Mapping the topography of the real-world landscape and its use in NICE

Real-world health, social care and public health data sources are numerous and previous reviews suggested that sources and access points were fragmented. A systematic search for all sources of real-world data was an undertaking beyond the scope of this project and would not directly address the research questions, which are focussed around the needs of NICE. Instead, three methods were used to map out the real-world data landscape with particular focus on those sources most relevant to NICE:

- A review of the literature on real-world data creating a map of relevant sources
- Semi-structured interviews with key stakeholders on their experiences of collecting or using real-world data and their broader perceptions on where opportunities for NICE lay
- Supplementary desk research including website searching

Results

Literature review & Interview Overview

Of the 548 studies published since 2005 and discovered through PubMed, 197 were excluded as they were either not on UK sources or were on veterinary real-world data (98), were not focussed on using or describing real-world data sources, provided a commentary about aspects of real-world data (e.g. ethics) without naming sources, or were duplicates. The remainder met inclusion criteria, including nine sources that were reviews or systematic reviews of different sources (most published close to 2005). Over a hundred sources (131) were coded as referring to ad hoc audits and data collection – these were studies that described conducting small surveys or clinical audits a single time and were representative only at a sub-regional geographic level; these are not named in this report as they are unlikely to be updated and are less generalisable in terms of geography, hence less relevant for NICE. The data sources named in the remaining studies were coded with simple descriptors (Appendix 2, analysis presented below) and were supplemented by those named by interviewees. Interviewees discussed a total of 55 sources of data that are condensed under 31 headings below, with additional information provided on each[5].

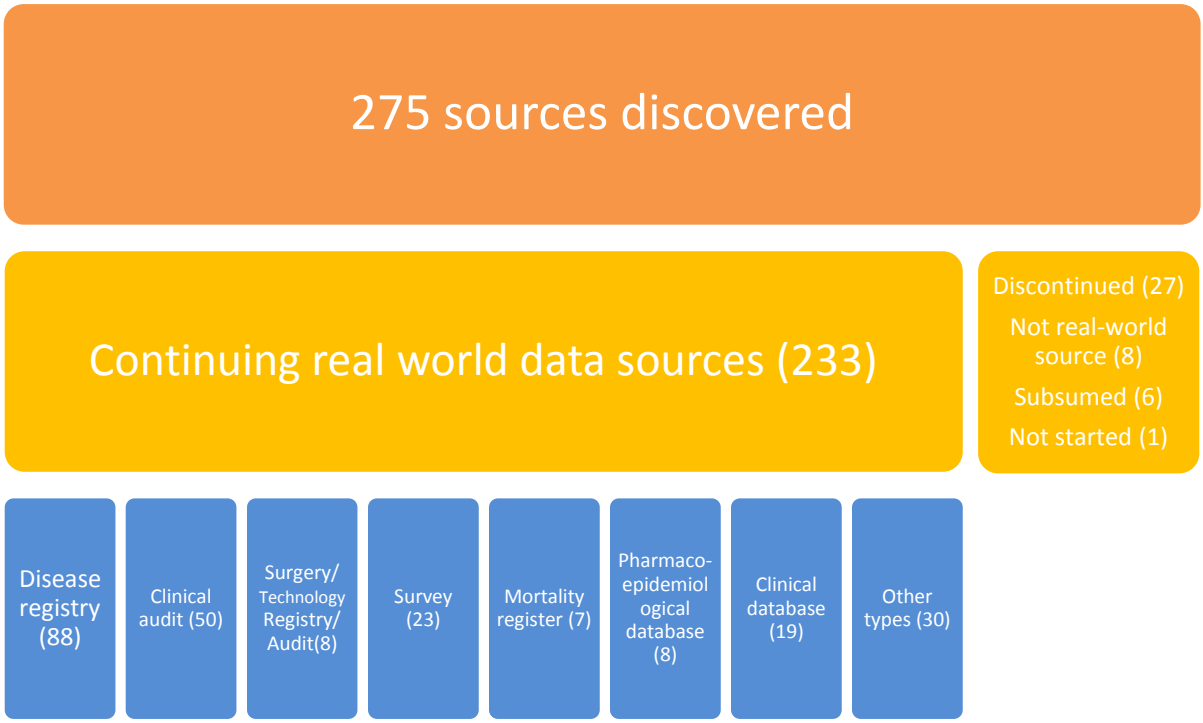
Mapping of real-world datasets

In creating the map based on the literature and on interviewees' responses we discovered a total of 275 different sources of real-world data (figure 1), of which 233 are analysed further, being of most relevance to NICE. The remaining data sources were found to either have been discontinued (26) or subsumed into other studies (6 sources), were not actually

[5] Note, we are still consolidating the results of the interviews and literature review

real-world data sources (e.g. they were procedures or standards for application in the real world (8 sources)), or were at the protocol stage (one source).

Figure 3: Sources of real world data discovered[6]



Sources were classified based on the name or description given on their websites, or as described in the literature. Disease registries comprised the greatest number of real-world sources uncovered (over a third); local geographic registries are represented as independent sources as they may have different specialisms/foci besides the geographic area. Clinical audits were the next more frequent type of real-world data comprising over a fifth of the sources uncovered (50 sources). Clinical databases and Health and Social Care surveys accounted for 19 and 23 sources respectively; a fuller breakdown is given below.

[6] It was also acknowledged that the HSCIC website held a great number of sources that could also be potentially profiled – future exercises could include a more detailed inventory of the HSCIC datasets.

Table 2: Real world data sources discovered in mapping exercises by type

Type	Number	Type	Number
Disease registry	88	Donation registry	2
Clinical audit	50	Social care database	2
Clinical database	19	Biomarker	1
Survey	23	Birth register	1
Pharmacoepidemiological database	8	Clinical database/ Disease registry	1
Surgery/technology register	8	GP Population List	1
Mortality registry	7	Incident registry	1
Screening register	5	Patient reported outcome database	1
Census	4	Population data	1
Workforce registry	3	Other	7

Real-world data on cancer were most frequent in our map of real-world sources (28 sources) followed by those described as 'general' health (18 sources). The latter included multidisciplinary studies as well as health studies that were not focussed on a particular discipline or experience, and most tended to be survey based (cross-sectional and longitudinal); data suitable for monitoring public health trends were also well represented among these survey data. An array of different health topics were represented, as shown in Figure 4 below. Figure 5, below, shows that close to half of studies covered the UK or Great Britain in their scope; and in total 143 sources were representative of England. In contrast, less than one-in-ten of the sources had local (sub-regional) coverage (reflective of our selection criteria)[7].

[7] Although this may also reflect restrictions imposed around ad hoc real world data sources.

Figure 4: Real world data sources discovered in mapping exercises by type

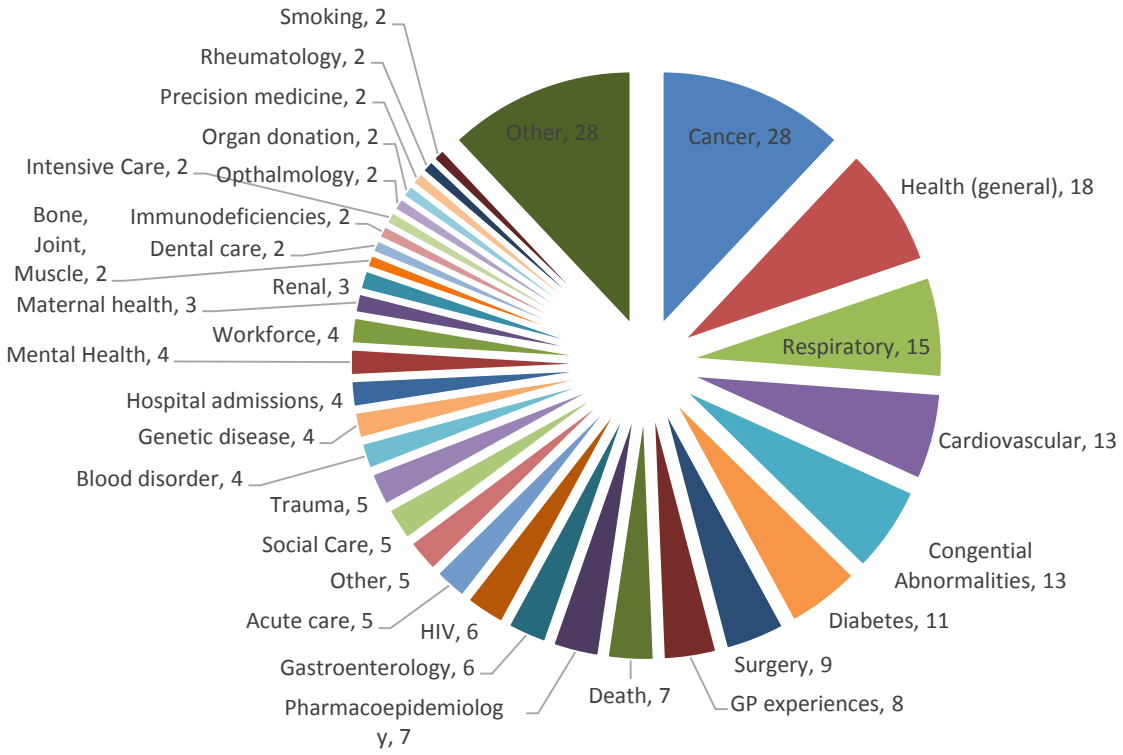
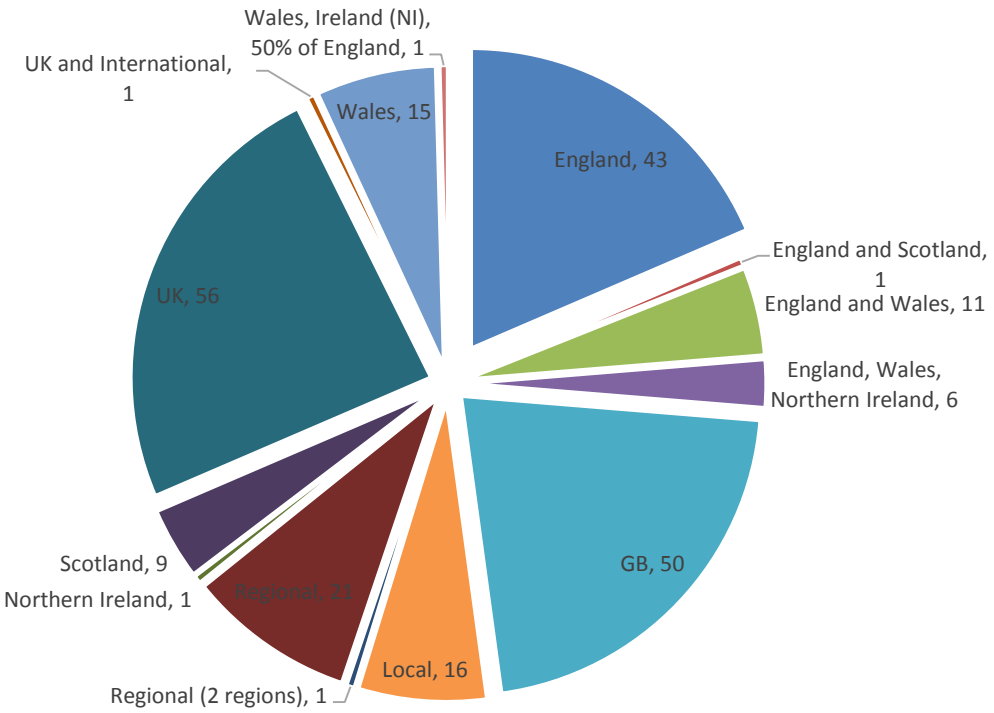


Figure 5: Real world data sources discovered in mapping exercises by geographic scope



An expert-driven map of real-world data opportunities

Deriving an expert map of real-world data sources and opportunities is not without its limitations and caveats, and the responses received are likely aligned with the experiences of the interviewees who agreed to participate, extensive as this was. In addition, naming of specific datasets for use by NICE without a specific research question is challenging. All interviewees were familiar with the work of NICE, and were given the background on how NICE envisaged greater use of real-world data in its work, and tended to focus their responses on datasets that they either knew or felt were either not utilised or underutilised. While the datasets named were in the context of all five aims of use of real world data for NICE, in practice most responses were actually focussed on the way in which NICE could measure the effectiveness of interventions or investigate epidemiological and social care trends.

Recommended sources to consider from interviews included:

1. **Understanding patient journeys and experiences using the broad scope of data contained within the National Cancer Data Repository**

Data collected and deposited on cancer and linked in the National Cancer Data Repository have been recommended for providing comprehensive data on a range of patient interventions and outcomes. These data have changed rapidly in recent years and the quality and breadth of the data have been transformed since 2008 to include core national outcomes data (Cancer Outcomes and Services Dataset (COSD)), radio therapy data (National Radiotherapy Dataset (RTDS)), cancer therapy data (Systemic Anti-Cancer Therapy Dataset (SACT)), and the National Cancer Patient Experience Survey; data from the ONS Minimum Cancer Dataset are also included to allow Repository Data to be reconciled with data from other sources. Three clinical audits also share data with the National Cancer Registries: the National Head and Neck Cancer Audit, the National Bowel Cancer Audit and the National Lung Cancer Audit. Efforts are made to link Repository data with other data including Hospital Episodes Statistics Data and data from the Clinical Practice Research Datalink, and linked data have been used to address a number of research questions aligned with NICE's intended use of real-world data. National Cancer Data Repository data have been used in a number of studies and are one of the most comprehensive sources of data on cancer available. Repository data now falls within the remit of Public Health England; some more recent publications, including the 2014 Cancer Patient Experience Survey, have not been updated on the NCIN website. The recommendation around the use of National Cancer Data Repository data is also accompanied by an offer to help understand the data and potentially work with NICE to use these data (details of which have been shared).

2. **Opportunities are available for assessing individual level service user outcomes through the Adult Social Care Survey (ASCS)**

The ASCS is designed to capture information on social care outcomes for Local Authority-funded social care service users, and provides unparalleled insights into social care user reported outcomes. Data from the survey feed into the Adult Social Care Outcomes Framework. The ASCS includes the items that comprise a preference-weighted measure of social care-related quality of life (SCRQoL) that form the Adult Social Care Outcomes

Toolkit (ASCOT). Apart from the ASCOT, the ASCS usually includes questions about socio-demographics; satisfaction with social care services; physical and mental health status; and contextual information such as accessibility of information and advice about support, services and benefits; and physical accessibility [34 58]. The survey is fielded as a cross-sectional survey and results are published through the Health and Social Care Information Centre; the geographical granularity of these published results reaches the Local Authority level only. The individual user-level data are not made available through sources such as the UK Data Service, unlike other user experience surveys such as the Community Mental Health Service User Survey, although user-level data have been used in research in the past (based on the existence of a small number of studies; for example [58]). ASCS was recommended on the basis of being one of the only large sources of data on social care user reported outcomes.

3. **Assessing resource usage using National Minimum Data Set for Social Care (Skills for Care)**

The National Minimum Data Set for Social Care (NMDS-SC) offers an insight into the characteristics of the social care workforce. It currently holds information on the workforce of approximately 25,000 institutions – Local Authority run as well as private institutions – and records on approximately 700,000 current or former social care workers are held [59]. The data include providers who are registered with the Care Quality Commission as well as those who are not required to register, such as providers of day care services [14]. These data were suggested as being potentially useful in addressing issues around resource usage and particularly in terms of identifying geographic differences in training and specialisms of the social care workforce. They have been used in a number of different ways including examining the characteristics of dementia care providers [14] and the contribution of migrant workers in providing social care [60]. The data cover England only, are published in aggregate form on the Skills for Care website, and are also available for research purposes from Skills for Care[8]. Data are based on establishment returns and include information on the care workforce in terms of: establishment details qualifications and training, job roles, demographic characteristics (gender, age, ethnicity), employment details, nationality, experience, sickness, pay, recruitment and retention.

4. **Exploring primary care practice using three of the large GP datasets**

Interviewees named three main sources of GP-based real-world data that could help to address NICE's real-world data needs. None of the interviewees prioritised any one source above another in terms of data quality, accessibility or cost; there nevertheless may be some differences in these characteristics as well as differences in terms of data linkage and size/coverage. We describe three of the sources mentioned below. All are based on the voluntary participation of GP surgeries; while the generalisability of these data sources is thought to be good, some studies have found that patients in some of these databases are more likely to be older and to live in more affluent areas than the population at large [61]. Each are widely used in the literature, generating over 150 publications combined annually [62].

[8] https://www.nmds-sc-online.org.uk/content/view.aspx?id=Accessing%20NMDS-SC%20data

4a. QResearch

QResearch is a large database of GP records from around 950 GP surgeries [63] from across the UK. QResearch was established in 2004 [64], later than other sources of GP records, although historical records in QResearch date back to the 1990s, making data from QResearch ideal for monitoring medical practice trends over time. QResearch data have also been used as the basis of multivariate prediction models to calculate risk of undiagnosed conditions among patients, such as the QCancer models [65]. Across the literature, the data are used for a variety of purposes aligned to NICE's intended use of real-world data. QResearch data are linked to deprivation data (based on patient postcodes) and cause of death data, and other linkage projects are underway in terms of linkages with registry and hospital episodes data [66]; in addition QResearch data have been combined with European records in drug safety monitoring research [67]. QResearch is a not-for-profit organisation established primarily to meet the research needs of academic researchers. Costs are associated with accessing the data dependent on size and focus of the data, and in addition to research teams being primarily based at academic institutions, at least one member of the research team must be a medically qualified academic registered with the General Medical Council to access data. *Note - QResearch data cannot be used for political purposes or for research that will not be published.*

4b. Clinical Practice Research Datalink (CPRD)

The CPRD is the successor of the long-established General Practice Research Database (GPRD), the GPRD being established during the 1980s as a patient health information system for use in General Practice[47]. It holds the records of over 13.7 million patients (representing over 64 million person-years of analysis [47]); estimates place the current number of patients currently alive at around 4.4 million and registered from 685 primary care practices spread throughout the UK. CPRD includes a combination of coded data, including on patient characteristics (some demographic information), clinical history, diagnoses, signs and symptoms, prescribing of drugs and devices, test results, referrals to secondary care and non-practice-based primary care services, immunisations and lifestyle factors. Free text data is also contained including the GP's notes and annotations, as well as a number of communications between the GP and other providers including letters and emails sent to and from the GP; some of these communications are even recorded as scanned images[47]. By 2011, data from the GPRD had appeared in over 850 peer-reviewed papers [47]; the CPRD has continued in this tradition and appears in a substantial number of publications. Its validity has also been tested for some conditions, being found to have high validity in some cases such as COPD [68] as well as many forms of chronic diseases [69]. CPRD data are the subject of several ad-hoc data linkage projects.

4c. The Health Improvement Network (THIN)

The Health Improvement Network (THIN) holds data from 587 GP surgeries representative of the UK population, who together hold the records from over 3.6 million active patients (amounting to over 85 million patient observed years). THIN represents a collaboration between IMS and a software development company (In Practice Systems (INPS)); in addition a dedicated unit exists within UCL expert in using THIN data. THIN data collection started in 2003, although the data collectors have a more extensive experience of working with primary care data pre-dating THIN [70]. THIN holds data on the patient (socio-

demographic information), the patient's medical history, practice-prescribed therapy, additional background health information such as smoking status, consultation episode data, and postcode-linked area characteristics. The data have been used extensively for purposes aligned with NICE's real-world data needs, including in studies monitoring the impact of implementing guidance and quality standards (for example in the care given to smokers [71]). The extent of data linkages beyond linkages based on patient postcode are not clear, although a number of studies have used THIN data linked with Hospital Episodes Statistics. As with other primary care databases, there is a cost in obtaining the data although support in assessing the feasibility of studies is free.

5. **Understanding the contribution of risk factors to disease outcomes using the Whitehall II study**

The Whitehall II study was named as a potential source of data for epidemiological research questions. The study recruited over 10,000 civil servants aged 35-55 working in London government departments in 1985. Since then there have been 11 sweeps of data collection with a third in progress. The study is also known as the Health and Stress study and has a focus on health and wellbeing that is not replicated across all the UK cohort studies. The narrow study design does hamper the generalisability of study findings to wider populations and would likely render it unsuitable for establishing prevalence or incidence rates representative of the wider population. However, the study has nevertheless been the basis for establishing associations between numerous exposures and patterns of health and ill-health, including health inequalities [72], and more recently in understanding predictors of onset of age-related diseases such dementia [73]. In addition to a breadth of sociodemographic and socioeconomic measures, the study also collects a wide array of biomarkers which have been used in some studies to establish the links between pharmacological interventions and disease outcomes, such as the impact of cholesterol lowering medicines and the relationships with physical exercise and diet [74]. The data are managed by a unit within UCL.

6. **Exploiting the longevity and near-universality of the Myocardial Ischaemia National Audit Project (MINAP)**

Cardiac audits were named by a number of our interviewees as a means of establishing the effectiveness of interventions (see also later entry). The Myocardial Ischaemia National Audit Project (MINAP) is a well-established audit with near-universal coverage of data on the management of heart attack across England and Wales (and Northern Ireland) since 1998. The audit focusses on the management of heart attack services provided by hospitals and ambulance services, and aims to provide a prospective 'census' of all patients receiving care, and benchmarks the care provided against nationally and internationally recognised timelines for care [75]. An extensive array of data are collected (130 fields) that capture elements of the entire patient pathway from the first call for help from the patient to the point of discharge. These data include demographic information, background medical history and clinical assessment, treatments, and a comprehensive array of information on drug therapy prior and during admission and at discharge. The audit is estimated to contain over 1.25 million records of patient experiences, collected by nurses and audit staff (often with the assistance of cardiologists) since it was established [75]. MINAP data are employed extensively for a

variety of different purposes [76], including for research purposes (since 2006) and studies have addressed a number of research questions including establishing the impact of different interventions [77], as well as epidemiological studies addressing questions around the antecedents of Mycardial Ischaemia [78].

7. Examining epidemiological trends using Health Survey for England

Health Survey for England (HSE) data were named as a source of data for examining epidemiological patterns. HSE data are available in summary form from the Health and Social Care Information Centre as well as individual records being made available for analyses from the UK Data Service. HSE is cross-sectional in design, with each wave including a different focus on a disease, condition or population group; for example past waves have included a focus on renal disease, wellbeing, sexual health, the health of mothers and children, the health of Black and Minority Ethnic people, and respiratory health. Whereas other many of the datasets named in this report require contact with service providers for the inclusion of individuals into the data, the HSE contains information on a full spectrum of the UK population. Some linkage has taken place between HSE data and other forms of data including Hospital Episode Statistics and National Cancer Registry data [79]. Sample sizes vary year-on-year, in part dependent on the study focus, but range between approximately 4,500 adult interviews in some sweeps to over 15,000 in others.

8. Examining life course experiences on patterns of ageing using the National Child Development Study

The National Child Development Study (NCDS) comprised a census of births occurring in a single week in 1958 and participants have been monitored into their 50s. The original 18,000 strong sample has reduced to a pool of around 9,000 active participants. A wide range of data have been collected relevant to different points of the life course. As the cohort ages, data collection has become more focussed on age related transitions, including cognitive decline [80]; the study has also recently started to collect biomarkers from study participants which have been used for genotyping studies [81]. In 2008, participants were asked for consent to link their data with NHS records with 79% giving consent [82], although to date, records have not been linked. Individual level data are available to researchers through the UK Data Service. The NCDS was suggested by an interviewee as a potential source of data on epidemiological trends among an ageing population. Other research questions may also be investigated, such as health or social care service usage.

9. Gaining a snapshot of social care and health service usage and needs of older people using the English Longitudinal Study of Ageing (ELSA)

ELSA is a longitudinal study focussed on older people aged 50 and over. The study originally recruited around 12,000 respondents with the first full wave of data collection occurring in 2002; since then the panel has been replenished three times to keep representation at younger age groups, so that in 2012 in the sixth wave of data collection, data were collected from 9,169 core study members. The study has been used to monitor

a number of health and social care outcomes and trends in England, as well as surveillance of risk factors, for example correlates of falls among older people [83]; some studies have also evaluated the impact of policy on older people's socioeconomic outcomes [84] and a similar scope may exist in terms of health and social care policy. Individual level data are available for research purposes from the UK Data Service. Data from the English Longitudinal Study of Ageing were recommended for use by NICE for investigation of epidemiological and social care trends among older people.

10. Understanding epidemiological and care trends among households, including ethnic minorities, using Understanding Society

Data from the UK Household Longitudinal Study, also known informally as the Understanding Society study can be used to examine a number of health and social care research questions. This study has now superseded the British Household Panel Survey as the UK's foremost source of longitudinal data at a household level [85]. It involves data collection from 40,000 households that have now been tracked annually over four waves of data collection from 2009/10- 2012/13 among the four constituent UK countries. Further information on the study design is found in the study documentation [86] and weights are provided by the study depositors to allow analyses of the data to be representative of the UK population. The large sample size, boosts of ethnic minority populations, and the wide breadth of the study content including detailed demographic, socioeconomic, health and wellbeing data allow researchers to investigate a number of different questions through a variety of disciplines. Epidemiological research using Understanding Society has included studies examining behavioural risk factors for obesity through patterns of diet and physical activity among young people [87] and the antecedents of hypertension among adults and the oldest old [88]. Data from Understanding Society were recommended for use by NICE for investigation of epidemiological and social care trends (within the context of older people, although the scope of the survey is broader than this).

11. Monitoring Social and Health care trends, experiences and monitoring the implementation of standards using Care Quality Commission data and reports

One interviewee, expert in social care data, described how Care Quality Commission (CQC) data and reports were a useful source in establishing trends and monitoring the implementation of guidance and standards. Another interviewee described how the CQC were "getting it – you don't have to use just this administrative data that is useless in measuring outcomes – there's all this data instead", and suggested the CQC as a case study of an organisation that was beginning to embed real-world data into daily practice. The CQC itself is dependent on returns, collects a number of different forms of data through inspection reports, and commissions a number of user experience surveys on different aspects of care, including clinical care experiences as well as community health and social care services. Survey data collected by CQC are generally made available through the UK Data Service for re-analysis.

12. Gaining an insight into patient experiences using Patients Like Me

'Patients Like Me' was recommended as a source of data for understanding the effectiveness of interventions through analysing patient experiences (as well as potentially providing data for economic modelling through quality of life measures). 'Patients Like

Me' provides data on patient experiences of treatments, as well more general data on symptomology, captured from web-based reports. Originally, the site restricted reports to a selected number of long-term conditions but since 2011 the site expanded to allow patients living with over 2000 different conditions to submit their data [9]. The self-selected nature of participation, its international scope, and non-standardised forms of data collection, do represent caveats to UK settings as well as in the broader potential of the data to address scientific research questions. Nevertheless, some clinicians are interested in the potential that data from sites 'Patients Like Me' can have on understanding the effectiveness of interventions, and more importantly in the impact that social networking can have on patient choice including in terms of adherence to treatment regimens and the impact on the patient-clinician relationships [89].

13. Tracking data on patient journeys in integrated delivery networks: the potential of Scottish Health Informatics (SHIP)

Data from Scottish Health Informatics Programme (SHIP) were named as being of interest for NICE in their potential capabilities to track patient journeys through different contact with health service providers and specifically within the context of integrated delivery networks. SHIP is a collaboration between academic partners and NHS Scotland, described as "an ambitious, Scotland-wide research platform for the collation, management, dissemination and analysis of Electronic Patient Records (EPRs)", while adhering to the Caldicott Principles around data collection [90]. This task is made easier as "compared with the rest of the UK, data quality is high and the centralisation of data in NHS Scotland is efficient"[10]. This is an ongoing research project, but a number of SHIP related studies have already been published, with many focussed on diabetes care [91 92].

14. Exploiting Hospital Episodes Statistics Data as a multipurpose dataset

Hospital Episodes Statistics (HES) includes data on 1.25 million hospital admitted patient, outpatient and accident and emergency visits annually[11]. The data are thought to be highly accurate and representative given that they form the basis for payment by the NHS to hospitals for the care they deliver. HES is an example of a clinical database whose primary function as an administrative tool has morphed to allow for research into health trends. Time series data can be constructed to allow examination of changing trends since the late 1980s. HES is routinely linked to ONS mortality data but is also linked to other datasets including CPRD and some registry data. Access to HES is obtained through specific requests from the Health and Social Care Information Centre, although some interviewees described access procedures as laborious. Some of the fields collected in HES include: Admissions; Period of Care; Augmented/critical care; Clinical data; Diagnosis; Discharges; Episodes and spells; Geographical data; Healthcare resource groups (HRG) data; Maternity; Organisation data; Patient characteristic data; Patient Pathway data; Practitioner characteristics; Psychiatric care data; Patient Socio-economic Data; and System data. HES data are already in use by NICE, although there may be greater potential to use these data in the five core intended uses of real-world data, particularly where data are linked. HES

[9] http://www.patientslikeme.com/about
[10] http://www.scot-ship.ac.uk/overview.html
[11] http://www.hscic.gov.uk/hes

data were discussed by interviewees as being suitable for use for a number of functions for which NICE wish to use real-world data.

15. Exploring Epidemiological Trends using the Avon Longitudinal Study of Parents and Children (ALSPAC)

The ALSPAC is a cohort study of women who were pregnant and with a due date between April 1991 and December 1992 living in the historic county of Avon. In total, over 13,761 women were recruited into the study contributing information on 13,867 pregnancies [93]; additional information from other eligible children has meant that data from a total of 15,247 pregnancies have been included in some form [94]. Information from mothers and children has been collected on a range of topics since then through questionnaires and clinical assessments; data have been linked with Mortality and Cancer registry data as well as CPRD and HES data [95]; and by March 2014 information from ALSPAC has appeared in over 1,000 peer-reviewed publications. ALSPAC data were suggested in terms of exploring epidemiological trends; the information may also have broader uses for NICE.

16. Using information from an integrated learning system through the Salford Integrated Record

One of the most consistent themes identified by interviewees was the need to expand on the data linkage properties of real-world data in order to track patient journeys through healthcare systems and ultimately improve patient outcomes. One example of a data source established with this aim is the Salford Integrated Record (SIR) which combines primary and secondary care contacts for around 97% of the population in Salford [9612], representing a population of approximately 300,000. SIR data are also extracted into the North West EHealth linked database (NWEH-LDB) and alert system, which was recently assessed in terms of its properties as a platform to support the delivery of pragmatic clinical trials [97]. *NB: This is one of a range of novel real-world data projects being undertaken in the North West (see http://www.herc.ac.uk/research-development/)*

17. Investigating the future application of GP records

Many interviewees named different sources of GP level data that is made available for research purposes (see above). However, GP data are inputted and stored on a number platforms and are interrogated through different software and queries – Apollo, EMIS, System1, Vision, MIQUEST – and these were also variously named by interviewees in different contexts. Not all the data collected through these systems are currently made available for research purposes, although the same software are used for other audit purposes (see HSCIC website), and further investigation may be required to investigate their current and potential future applications.

18. Understanding the effectiveness of interventions using National Joint Registry as a registry that is collecting longitudinal outcomes and patient reported outcomes

The National Joint Registry was established in 2002 to collect information on all hip, knee, ankle, elbow and shoulder replacement operations and to monitor the performance of

[12] http://www.salfordccg.nhs.uk/documents/Publications/SIRA5Booklet.pdf

joint replacement implants[13]. It covers all replacement operations conducted in England, Wales and Northern Ireland and is one of the few disease registries to collect data on patient reported outcomes including quality of life, and also measures some indicators longitudinally (six months after the procedure). The registry has an active research programme and data from the registry are used to explore a number of biological, mechanical, clinical, economic, and social factors influencing the outcome of joint replacement. Data from this registry are recommended for use by NICE in investigating the effectiveness of interventions in real-world settings.

19. Understanding the effectiveness of interventions and monitoring the impact of guidance using the Sentinel Stroke National Audit Programme (SSNAP)

Different interviewees recommended data from the Sentinel Stroke National Audit Programme for use by NICE for providing a comprehensive understanding of the effectiveness of stroke care. These data were recommended as one of the few audit datasets that collect data longitudinally, although in the case of SSNAP this is a recent development. SSNAP aims to collect a minimum dataset for every stroke patient, including acute care, rehabilitation, 6-month follow-up, and outcome measures in England, Wales and Northern Ireland[14]. A key feature of the SSNAP programme is commitment to 'harness the power of "Big Data" to produce near real-time data collection, analysis and reporting' [98]. SSNAP builds on a previous 15 year history of collecting information on stroke outcomes. SSNAP data have been used to monitor the effectiveness of interventions, monitor epidemiological trends as well as monitor the implementation of stroke guidance.

20. Understanding the effectiveness of interventions and monitoring the impact of guidance using the Renal Registry

Data from the UK Renal Registry (UKRR) were recommended on the basis of having 'huge amounts of data' from one interviewee, and were viewed as having potential in fulfilling many of the ambitions NICE has for the use of real-world data. The UKRR itself views itself as being one of the few high quality registries that is open to requests from researchers. The UKRR collects data from 71 adult and 13 paediatric renal centres across the UK. In 2014, it published its 17th report that included demographic, biochemical, treatment and therapy characteristics, transplant waiting list information, geographic and demographic inequalities in access to services, as well as a number of key outcomes for renal patients[15].

21. Understanding the effectiveness of interventions and monitoring epidemiological trends using Adult critical care case mix programme (managed by ICNARC)

ICNARC (Intensive Care National Audit & Research Centre) run the Case Mix Programme (CMP), an audit of patient outcomes from adult, general critical care units (intensive care and combined intensive care/high dependency units) with near-universal coverage (99% of critical care units) and collecting units from England, Wales and Northern Ireland.

[13] http://www.njrcentre.org.uk/njrcentre/default.aspx
[14] https://www.rcplondon.ac.uk/projects/sentinel-stroke-national-audit-programme
[15] https://www.renalreg.org/wp-content/uploads/2014/12/Report2014.pdf

Currently, the data collectors estimate that the data from 1.5 million patients are held[16]. Data quality is likely to be high, as reports are subject to 600 validation checks. Data collected for the CMP include unit identifiers, demographics (e.g. age, sex, and ethnicity), case mix (e.g. acute severity, comorbidity, surgical status, reason for admission), outcome (e.g. unit/acute hospital survival) and activity (e.g. unit/acute hospital length of stay) for each admission to each critical care unit[17]. The data have been used to address a number of research questions around epidemiological trends, such as the antecedent characteristics for sepsis for example [99], and around the effectiveness of interventions and examining the outcomes of critical care patients for example [100].

22. Harnessing the potential of cardiovascular audit and register data to address NICE's real world data needs

A number of interviewees gave linked responses around the potential of a number of different sources of real-world data on cardiovascular disease and treatment. These tend to fall within the remit of the National Institute of Cardiovascular Outcomes Research Unit (NICOR) who run six National Clinical Audits (including MINAP, which was specifically named and described earlier) and five Healthcare Technology Registers, which record the implementation of cardiac related devices. The data tend to have good geographical coverage across the UK. NICOR welcome applications and collaborations for research using the data and also supply information to the DH and the Care Quality Commission. NICOR are experienced in data linkage projects with Hospital Episode Statistics (HES) data and Clinical Practice Research Datalink data. The six clinical audits (commissioned by the Health Quality Improvement Partnership (HQIP) include:

a. Adult Cardiac Surgery Audit (with the Society for Cardiothoracic Surgeons) – collecting all data on major heart surgery in the UK
b. Adult percutaneous interventions audit – collecting information on adults receiving percutaneous cardiovascular intervention
c. Cardiac Rhythm Management Audit – collecting information on adults receiving interventions for cardiac rhythm disorders
d. Congenital Heart Disorder Audit – collecting information on children who receive intervention
e. Heart failure Audit – collecting information on emergency admissions of patients with heart failure and their outcomes
f. MINAP

The five technology registers include:

(a) Left atrial appendage occlusion register – monitoring the effectiveness of this intervention in preventing stroke
(b) UK Neuromodulation registry – collecting data on all spinal cord stimulator (SCS), intrathecal drug delivery implant for pain and spasticity (ITDD) and peripheral nerve stimulator (PNS) procedures
(c) Percutaneous mitral valve leaflet repair for mitral regurgitation (register) - evaluating percutaneous mitral valve repair effectiveness

[16] https://www.icnarc.org/Our-Audit/Audits/Cmp/About
[17] https://www.icnarc.org/Our-Audit/Audits/Cmp/About

(d) Patent foramen ovale closure – monitoring the effectiveness of this intervention in preventing stroke
(e) Transcatheter Aortic Valve Implantation (TAVI) (register) – measuring the characteristics and clinical outcomes of patients receiving TAVI

NICOR data is widely used in the literature for all of the functions for which NICE wants to use real-world data. NICOR audit data were recommended on the basis of the comprehensiveness in terms of scope, their near-universal nature, and their potential for research which aligned with NICE's ambitions around real-world data.

The interviews also revealed that work was being undertaken at the Farr Institute to link cardiovascular disease data (MINAP and other sources) with CPRD data.

All information featured on the NICOR and/or from interviews: http://www.ucl.ac.uk/nicor

23. Utilising data from the National Diabetes Audit; "the most advanced for long-term conditions"

The National Diabetes Audit was described by one interviewee as being 'the most advanced [source of audit data] for long-term conditions' due to its breadth and longevity. It is described by the Health and Social Care Information Centre as being the largest clinical audit in the world[18], and integrates data from primary and secondary care providers, and sets out to address four key questions revolving around the diagnosis, care and treatment of complications among diabetes patients in England and Wales. The audit is commissioned by HQIP and conducted by HSCIC and Diabetes UK. It includes (i) National Diabetes Core Audit (NDA); (ii) National Diabetes Inpatient Audit (NaDIA); (iii) National Pregnancy in Diabetes (NPID) Audit; (iv) Patient Experience of Diabetes Services Survey (PEDS); (v) National Diabetes Audit of Footcare (NDFA). Data from the audit are used for a variety of research purposes and are likely to be useful for all the functions for which NICE wants to use real-world data.

24. Capturing genetic information on biomarkers in the UK Biobank

One of our interviewees expressed the view that biological markers needed to be promoted in epidemiological research and cited the UK Biobank as a development in this field.

25. Calculating cost effectiveness based on data from the Personal and Social Services Research Unit

The Personal and Social Services Research Unit (PSSRU) calculate unit costs for different social care packages in a unit cost report published annually. One interviewee recommended these data as the basis for cost effectiveness studies. The PSSRU also has a much broader remit and scope, being responsible for the development of many advances in social care data collection and measurement instruments including the ASCOT measurement tool for social care-related quality of life for individuals.

[18] http://www.hscic.gov.uk/nda

26. Mental Health and Learning Disabilities Data Set

The Mental Health and Learning Disabilities Data Set (MHLDDS) data were described as one of the few real-world data sources capturing data on mental health by one interviewee. Formerly known as the Mental Health Minimum Data Set, these capture data about care delivered to users of NHS funded secondary mental health and learning disabilities services for adults in England[19]. While the scope of the data are relatively narrow, focussing on service usage rather than outcomes, the data may nevertheless be useful in establishing broad trends and standards around mental health; for example the 2014 focus on the distance travelled to access mental health services may help develop and monitor the implementation of quality standards around accessibility.

27. Understanding trends in screening rates, healthcare and epidemiology using Quality and Outcomes Framework (QoF) data

QoF data have an advantage over the primary care databases named above in having much higher levels of GP surgery coverage than all three combined (over 7300 in 2013/14). QoF data are based on an incentive programme open to GP surgeries in England that measures how GP surgeries are performing against indicators in terms of their clinical performance; public health responsibilities (a core set and an additional set); quality and productivity; and patient experience (other factors taken into consideration also include surgery workload, local demographics and the prevalence of chronic conditions). Individual patient-level data however are not collected which impede the usefulness of the data for NICE's intended uses of real-world data. Nevertheless, the data may provide indicative evidence around questions relating to the monitoring of epidemiological trends and understanding the impact of implementing NICE guidance through monitoring trends over time at the surgery level.

28. Understanding epidemiological trends and measuring the effectiveness of interventions using the UK Inflammatory Bowel Disease Audit

The UK Inflammatory Bowel Disease Audit was discussed by one interviewee in the context of measuring an improvement in the treatment provided to patients (suffering with bowel disease) and being able to examining 'cycles of improvement'. It is managed by the Royal College of Physicians. In the fourth round, reporting in 2014, the audit was composed of five elements (i) inpatient care audit (collecting information on the first 50 patients admitted with ulcerative colitis in each hospital); (ii) inpatient experience questionnaire (each patient included in the inpatient audit provided with a satisfaction questionnaire); (iii) biological therapy audit (treatment, delivery, disease activity and quality of life in patients who are prescribed Infliximab or Adalimumab for Inflammatory Bowel Disease); (iv) an organisational audit; (v) quality improvement initiative (sharing best practice). Data from the audit have been used for a variety of purposes of relevance to NICE, including examining resource use and exploring staffing requirements in IBD centres [101].

29. Harnessing the potential of audit data to address NICE's real world data needs through clinical audits conducted by the Royal College of Surgeons Clinical Effectiveness Unit

[19] http://www.hscic.gov.uk/catalogue/PUB16421

The Royal College of Surgeons (RCS) Clinical Effectiveness Unit (CEU) was named as a source of audit data that may address NICE's real-world data needs across a range of conditions. This source of real-world data was named due to the breadth of audits being conducted under the direction of a single organisation, which may facilitate access and relationships and promote uniformity in data collection methods. The audits conducted by the RCS-CEU include:

- National Audit of Oesophago-gastric Cancer – monitoring the care of patients with Oesophago-gastric Cancer.
- National Bowel Cancer Audit - A national clinical audit monitoring the care of bowel cancer patients.
- National Prostate Cancer Audit (NPCA) – Measuring standards around the care that men receive following a diagnosis of prostate cancer.
- National Vascular Registry - Collects data and reports on the process of care and outcomes following: (i) Repair of abdominal aortic aneurysm; (ii) Carotid interventions to prevent stroke; (iii) Operations of the lower limb, including angioplasty, infrainguinal bypass and amputation.
- CRANE Database – A registry of all children born with cleft lips and palates in England, Wales and Northern Ireland
- National Emergency Laparotomy Audit (NELA) – An audit of emergency abdominal surgery in England and Wales

*All data taken from the Royal College of Surgeons website https://www.rcseng.ac.uk/surgeons/research/surgical-research/ceu/projects

30. Data used to populate NICE's Return on Investment Tools

Several sources of data are used to populate NICE's return on investment (ROI) tools that aim to inform decisions taken by Local Authorities as to the likely impacts of their commissioning decisions. For example, the NICE ROI tool to understand the impact of tobacco use strategies and interventions uses data from several nationally representative data sources as the basis for estimates used in the tools – including the Integrated Household Survey (IHS), General Lifestyle Survey, statistics from the National Centre for Smoking Cessation and Training, and statistics from Annual survey of hours and earnings among others [102]. These were suggested as sources of public health real-world data.

31. Data from private health providers

Data from private health providers (for example Bupa) were named as a potential source of data where the continuity of care and treatment pathways could be examined. While these data would have the disadvantage of being based on a small, unrepresentative (socioeconomically advantaged) section of the population, they may provide some of the only real-world data sources that could be used to monitor the continuity of care between clinical care and social care. Issues of access to these data and the quality of these data

are unknown. However, these data were identified as overcoming the barrier around *'NHS databases where patients become invisible when they move from one setting to another'*.

Overall summary of real-world sources of data recommended:

Themes: A number of themes emerged around the data sources recommended, which are also explored in section 4. These include:

- The importance of regarding clinical audit and disease registry data as important and overlooked forms of real-world data
- The importance of regarding survey data as an important and overlooked form of real-world data, particularly in terms of social care (and public health)

Few datasets were mentioned twice (with exception of Stroke and cardiovascular audit data); in contrast there was greater saturation in the themes emerging around the properties and challenges of working with real-world data

Gaps in real-world data sources: Some gaps have emerged in the map of real-world data:

- An absence of real-world mental health data. This was recognised by one interviewee who observed *"that in Cinderella area such as mental health, we have not tended to do a lot of measurements; [we need to] get them up and running"*.
    - From the additional mapping exercises an additional five sources of relevant mental health data were identified, of which the 'Community Mental Health User Survey' and the 'Prescribing Observatory for Mental Health' may be particularly relevant in addressing NICE's real-world data needs
- Data specifically focussed around resource use
    - In addition to the social care workforce data identified above, two data sources on the GP workforce and Diabetes specialist workforce were identified from the additional mapping exercises.
    - From the additional mapping exercise, additional real-world data sources that provided pharmacoepidemiological information on medication usage were identified: Tayside Medicines Monitoring Unit (MeMo); Prescribing Analysis and Cost (PACT) data; Hospital Pharmacy Audit Index; Electronic Prescribing Analysis and Cost Tool (ePACT); Prescription Pricing Authority database; IMS Health databases (Medical Data Index); IMS Health databases (MIDAS Prescribing Insights); Prescribing Observatory for Mental Health; British Society of Rheumatology Biologics Register (BSRBR); UK HIV Drug Resistance Database; Biologics were also represented in the Bowel Disease Audit
- Data sources that incorporated patient reported outcomes and digital technologies as standard were in the minority; although there are efforts underway to improve the recording of Patient Reported Outcomes taking place across the NHS[29]
- Few specific data sources from commercial delivery organisations such as data from private healthcare insurers were identified; no data sources on over-the-counter medications were identified; few data sources from the voluntary sector were identified except where organisations were working in partnership with statutory organisations or Royal Colleges (e.g. in the case of Diabetes UK) or where

their activities were linked to quality standards. One source did state that data from charities were often overlooked.
- Qualitative real-world data in health and social care virtually absent
- Finally, the direction of real-world data is to be able to link different sections of patient journeys as standard. Few models emerged that were actually able to perform this function as standard, a notable exception being the Salford Integrated Record.

How is NICE currently using real-world data?

NICE conducted a review of its internal use of real-world data across its Impact and Evaluation; Costing and Commissioning; Internal Clinical Guidelines (CCP); Health and Social Care; Safe staffing; National Clinical Guideline Centre, National Collaborating Centre for Mental Health, National Collaborating Centre for Cancer and Social care teams. This review asked teams to name which data source was currently being used, how these data were accessed, processes employed for accessing data, associated costs; and a brief description of how the data were used. From this review, twenty-four datasets were identified in use (with some having multiple uses – a total of 37[20] different combinations of datasets/uses/teams were reported across NICE; see Appendix 4 for full list):

- The majority of data are used, internally at least, to either (i) inform on the uptake of NICE guidance and/or explore use of medication (nine reports could be described in this way (one represented future plans)); or (ii) for health economic modelling (twenty reports could be described in this way).
- Some of the data appear to support the development of quality standards particularly around safe levels of staffing (three reports could be described in this way)
- There was one reported use of data for monitoring epidemiological and demographic trends (ONS data was reported as being used for this purpose, although the precise dataset was not reported); Lifetable and mortality data was also used
- One dataset was described as being used to establish the effectiveness of interventions (Primary Care level data using the IMS Disease Analyser; see [103] for an example of NICE's use of these data)
- There was a mixture between data that was accessed in aggregate form (pre-constructed tables) and patient/surgery/hospital level data that was used for further interrogation and re-analysis
- Clinical audit/disease registry data was reported twice: NICE is currently engaged in discussions to access the Systemic Anti-Cancer Therapy (SATC) dataset to gather data on the uptake of medicines; clinical audit data was identified as being used although the precise nature of use and the name of the audit was missing.
- Two forms of survey data were reported – the Health Survey for England and the National Cancer Patient Experience Survey; the former was reported as beneficial for quality of life data and data on ethnicity
- One Primary Care database was in current use – The Health Improvement Network data (on request from HSCIC); primary care data is also found through the IMS Disease Analyser

[20] Excludes one dataset that was not deemed to be real-world data

Section 2: (Findings I) Mapping the topography of the real-world landscape and its use in NICE - How is NICE currently using real-world data?

- One of the most frequently accessed datasets Hospital Episodes Statistics data
- Licences have been purchased for three datasets for internal analyses: IMS Disease Analyser data, Hospital Episodes Statistics and the UK Nursing Database
- The most common gateway for accessing health and social care data was through the Health and Social Care Information Centre

How does the way in which NICE utilises data differ from the topography of the real-world data landscape

A full assessment of the way in which NICE utilises real-world data is only be possible with the addition of input from external agencies and partners through which NICE may commission work. Furthermore, other uses for data may be supported under broad headings in Appendix 4 such as 'ad hoc' enquiries or 'ONS'. However, the evidence suggest NICE's internal use of real-world data differs from the real-world data landscape in the following ways:

- **The use of clinical audit data by NICE does not match the widespread availability of these data**; no specific clinical audit data were mentioned as being currently in use. In addition, one of our interviewees emphasised this through observing *"So it's National Clinical Audits that are hugely useful and underexploited for all the five legitimate purposes that absolutely NICE need to fulfil"*
- **The use of disease registry data by NICE does not match the widespread availability of these data**; no specific disease registry data were mentioned as being currently in use
- **Survey data are underutilised** and was mentioned explicitly on two occasions (although may be captured under broad headings such as 'ONS' or 'HSCIC' data; further explication was not provided on these)
- Most of the data sources/functions currently in use appears to allow for cross-sectional analyses or repeated cross-sectional analyses only; patient-level longitudinal analyses appear to be conducted rarely
- Few datasets used by NICE capture Patient Reported Outcomes; one exception is the National Cancer Patient Experience survey
- NICE use an extensive array of different datasets to support understanding trends and the economic modelling of changes in prescribing trends
- There were no reports of NICE requesting additional data to be collected alongside standard data in any of the real-world sources
- Few real world data were reported as being employed to research the effectiveness of interventions
- Few real world data were reported as being employed to establish epidemiological trends
- Primary care data is mainly accessed through sources based on The Health Improvement Network data.
- Few of the existing sources allow for linkage across different services which a patient or service user may experience; however, these types of data were also underrepresented in the results of the mapping exercise

- For social care, none of the datasets described are directly sourced and explicitly focus on monitoring trends in private provision despite private provision being hugely important for the sector; these data were also underrepresented in the results of the mapping exercise

In-depth choices for mapping

From this long-list of 30+ data sources, a selection of eleven was chosen for in-depth profiling based on input from a NICE steering group (see Section 5). Datasets were prioritised if they were those that were not in current use by NICE and where they appeared to meet some of the broader gaps in usage and the themes emerging from the interviews (see section 3). Although the long list of recommended datasets presented included a number of audit and registry data sources that NICE were not currently using, most of these were not selected for in-depth profiling because it was felt that there would be high levels of organisational familiarity with these datasets. Instead, an allied separate exercise was put into place to map out how NICE used registry and audit data and how this could be expanded in the future; the results of this exercise will be published separately from the present study.

We used the template developed in Section 1 around the characteristics that make for good quality sources of real-world data as well as assessing the utility of the different sources for NICE's intended purposes as the framework for profiling each data source. This framework was populated through a narrative review of the literature as well as reviewing any accompanying study documentation; some information on the data was also collected from the interviews. The exercise of profiling each data source was conducted in abstract without a fixed research question defined, and consequently we do not give an overall score or rating for any data source. Nevertheless, it is clear that all the data sources included for in-depth profiling could make substantial contributions to the work of NICE as is the case for those included in the 'long-list'. While the profiles give a grounding in some of the properties of the data, such is the complexity of all of the data sources, any use of the data by NICE will require significant investment in terms of time and training, and in some cases, cost. However, as is clear from the profiles, the returns on this investment are potentially large when beginning to use any one of the data sources included here, or any of those described earlier in this chapter. Section 4 includes the detailed profiles while section 5 includes some analysis of the findings of these.

The data sources selected for in-depth profiling were:

1. **QResearch**
2. **THIN**
3. **CPRD**
4. **ELSA data**
5. **NMDS-SC**
6. **Salford Integrated Record**
7. **Community Mental Health User Survey**
8. **Prescribing Observatory for Mental Health**
9. **Health Survey for England**
10. **Adult Social Care Survey**
11. **Care.data**

Section 3 (Findings II): Broader cross-cutting debates and themes

Cross-cutting themes

What does real-world data mean?

We include a discussion on the meaning of real-world data as this may help to shape the remit of any real-world data projects that NICE may undertake in future.

Most of our interviewees described the tenets outlined earlier as being indicative of real-world data and described real-world data as being '*about types of services, people that use the services and how the money is spent*'. This aligns closely with conceptualisations of Healthcare Technology Assessment itself in the literature, as the study of 'medical, social, ethical and economic implications of development, diffusion, and use of health technology' [104]. Such alignment reinforces the role of real-world data as an essential tool for NICE. In describing real world data, there was a tendency among some to focus on routinely collected/administrative data, particularly among clinical and public health experts, and an emphasis on the quality of the data: *"[real world data is] electronic pulling of data from lots of different sources with standard quality assurance processes - accepted in a huge range of formats and transformed. So very extreme: [at] one end hand written facts from doctor's letter to a completed formatted xml structured file that has all data and fits a national file… you can accept anything in-between algorithms can be put in place that flag the quality of data"*.

This aligns with the literature where much of the data collected in primary care is viewed as a by-product of administrative activity. Some experts had a broader conceptualisation, seeing real-world data as being data that *'provides insight on patients in the natural setting for that patient'*. Among social care experts, the definition of real world data tended to be broader from the outset and survey data was much more likely to be included in accounts when social care expert interviewees described different forms of real-world data. Social care experts stated that the terminology of 'real world data' was less common in the field, although an allied term was not used in its place. This may reflect the findings of previous studies where inconsistent terminology was viewed as a barrier to mapping the availability of real-world data holistically [3]. Some interviewees also stressed a distinction between real-world datasets which may represent amalgamations of multiple primary sources of data, and data sources, which were suggestive of primary or single sources.

The breadth in the definition of real-world data was viewed as problematic by some; for example one interviewee could foresee that bringing too much breadth into a definition of real-world data increased the risk of diluting the quality of data under consideration: *'Real world data or all world data or just all data'*. While most interviewees were comfortable with surveys and other forms of structured observational data being included in a definition of real-world data, newer potential sources of real-world data, particularly data that could be collected through apps or from unpublished sources, or views data from

other forms of social media, were not mentioned by interviewees. One exception was that one social care data expert described how blogs and other alternative sources of data, including Local Authority committee meeting minutes were useful sources for discovering the existence of other, less well known or not routinely published information, but not as primary sources in of themselves. Data from these sources will likely be of lower quality and subject to fewer robustness checks compared to larger data sources, although the content may nevertheless be unavailable from other sources.

Not all interviewees were familiar with, or agreed with the terminology 'real-world data', with one interviewee observing *"I don't see any unreal data"*. Some preferred referring to these data more transparently as 'routinely collected data'. This may be worth noting when NICE are engaging in discussions about these issues externally.

Tracking patient and service user journeys

Two intertwined recurrent themes emerged from the fieldwork conducted, and these are not explicitly highlighted in these terms in the literature. Interviewees were asked about data that were suitable for tracking patient and service user journeys and trajectories. Overwhelmingly, respondents across all sectors described that this undertaking was fraught with difficulty. Much of this challenge was attributed to difficulties in being able to link data between sources using a common identifier, although among those respondents who were more expert in clinical health sources it was recognised that considerable efforts were being undertaken to link across different sets.

> *"Two important datasets - and they just don't talk to each other…..We have all these datasets not linking at the minute and [those responsible] are blocking any linkage"*

A good example of the difficulty in tracking journeys was described by an interviewee expert in clinical health data sources. He described a situation where an individual had suffered a heart attack and was diabetic. Their records should appear across multiple sources of data – for example in Myocardial Ischaemia National Audit Project (MINAP) data, National Diabetes Audit data; they should also be recorded in Hospital Episodes Statistics Data and in Clinical Practice Research Datalink (CPRD) data. However, the record of the same patient having suffered a heart attack and being diabetic appearing consistently across different data sources was thought to be relatively low – our interviewee gave an estimate of this occurring two-thirds of the time (although stressed that further verification was needed around this estimate). Regardless of the specific level, in common with other interviewees he identified that consistency across data sources hampered obtaining an accurate record of patient and service user journeys and trajectories. Inconsistencies reflected non-standardised data collection practices which led to differences in constructs and the type of information collected, but also in terms of the timeliness and updating algorithms of the data sources.

> *"So I think there is a lot of data out there which would be good for multiple purposes not necessarily the primary purposes it has been gathered for. But equally there is a total lack of joining up."*

This theme was also present in the interviewees who spoke about social care data. Here the challenge was compounded by the issue of defining care packages and care trajectories, and the disparity between the inflexibility of data collection systems and the multiplicity of different care pathways and providers who would provide care: *"working with them [real world data] trying to work our social care issues how traditional 'care pathways' work - they are problematic in social care with the multiplicity based on the individual person care package-environment mix"*. However, one social care interviewee suggested that this picture was changing as they gained access to case records directly from local authorities, suggesting that social care real-world data access may be highly dependent on establishing good relationships with individual authorities:

> *"we are starting to use information from case records, this is in partnership with Local Authorities , and they are able to share some of that information for us to use"*

However, while most interviewees have been pessimistic about the capacity to trace patient and service user trajectories and journeys, some did express the view that this was a rapidly changing situation. For example one interviewee spoke of the rapid developments in linking different types of cancer data in forming the contents of the National Cancer Data Repository, and the rapidly expanding breadth of data being linked, including radiography and chemotherapy treatment and outcome data that enable the improved tracking of patient journeys. Some of these linkage exercises have taken place on a pilot basis where only extracts of data have been linked whereas others are now fully operational. Many of these developments had only taken place in the past five years, and clearly this is a rapidly changing field. In addition, a common theme among some of the interviewees was the importance of user and data community groups in helping to progress the breadth and usability of different datasets. Representation on such user groups can enable organisations and individuals to both keep pace with developments but also to influence how sources of real world data are enhanced to meet users' needs. Nevertheless, despite these positive glimmers, the overwhelming theme in these findings is that tracing patient and service user trajectories and long-term outcomes is challenging, and where this is possible, it likely means working closely with data depositors and controllers to shape and customise the dataset. An example of such a collaborative approach is found in the Salford Integrated Record project.

This theme does complement previous literature in this area, but is expressed differently in the literature through emphasis on the substantial fragmentation in sources of real-world data [2 3 18 19 105], which presumably hampers tracking patients and service user journeys. For example Newton and Garner [3] had discovered up to 400 specific clinical

registers in England alone in their review; Raftery and colleagues discovered over 270 sources of real world data in their comprehensive review of registries and databases [19]; more recently Rankin and Best [105] identified approximately 20 registers with a narrow focus on four childhood chronic diseases. There are likely to be considerable overlaps between such fragmented sources of data. In our interviews the fragmentation in data sources was expressed in terms of the consequences of this fragmentation, and in particular the difficulty in tracking patient and service user journeys longitudinally. The absence of a common identifier between sources was identified as a prime contributory factor to this.

> "*if you want to understand hospital admission for example in relation to social care episodes and you want to identify the same people who enter with respite care then go out for hospital operation then to an agency for domiciliary care you would need to have key I.D for this and these are not available – we have leads in terms of the health side but not the social care side*".

However, new data sources are being developed and many of these are based on the linkage of other routinely collected health and social care data sources, including those intended to track healthcare pathways and contact with different providers (for example the Suicide Information Database-Cymru [42]). Elsewhere technology and the use of routine data (for example through electronic health records) have been identified as critical in meeting new challenges, such as implementing integrated care systems [46]. The literature and our interviews confirm that while historically there have been poor mechanisms for tracing patient journeys, this is an evolving picture. The question is whether this situation is evolving quickly enough to meet the demands of the changing health and social care landscape, and particularly in terms of providing evidence to support these big challenges.

Determining treatment options and quality of data

Real world data's particular strength is the potential to provide the most complete picture available about the health and care patterns of the nation as conceivably possible [2]. This picture is composed of data that isn't derived from an unrepresentative subset of the population, but provides a population-based snapshot of contacts with providers that take place and information on illness or care needs, treatments or care packages, and ideally provides enough information on the outcomes of individuals and their use of other health and social care services. The quality of each data set varies greatly in content and what type of data has been collected. It is therefore up to the scrutiny of individual and what is required from the data as to where the balance lies between breadth and depth.

> *"There is a trade-off between having more information in terms of numbers and information in terms of breadth and depth of indicators. So survey data such as ELSA [English Longitudinal Study of Ageing] will give you a lot more in terms of quality of*

characteristics – income, wealth, needs, households' composition, service users etc. Certain outcomes will be much more limited on the other hand data from services; there will be thousands of cases in other sources - but much more limited – and the data and may not be of the same quality. We try to combine the data, look at patterns from both".

Data quality has been recognised as an issue of real-world data by many in the literature [1] and despite the population-based design, several groups are known to be underrepresented in these data. These include women, children, the very elderly, ethnic minorities and those with multiple co-morbidities [37]. Some studies suggest that where patients are missed in real-world data sources, they may have systematically different characteristics. In one examination of the MASCARA (Manejo del Síndrome Coronario Agudo Registro Actualizado) register, a hospital-based cardiovascular register, patients who were 'missed' off the register had higher risk profiles of cardiac events and received fewer recommended therapies than included patients; in addition mortality was almost three times higher [40]. Furthermore, the study also found that few users of disease registry data acknowledged the potential selection bias in the limitations of their findings [40].

Representativeness of data and information from hard-to-reach groups may be improved through obtaining multi-site data, although this can introduce further problems around fragmentation and diversity in administrative systems.

Our interviewees described problems in the quality of the data around these groups and more often pointed to their invisibility in several real-world data sources (i.e. several sources do not collect information on ethnicity, sexual orientation etc.), impeding the ability to research inequalities in the services and treatments that different groups receive. More importantly numerous other fields around co-morbidities and the receipt of other treatment regimens may also be missing [34]. Some interviewees described approaching or working with other organisations to obtain real-world data where possible; for example they described approaching bodies such as the Equalities and Human Rights Commission or charity or representative groups such as Stonewall. However, these data may have a different set of strengths and weaknesses attached to them and in particular, such data may not be subject to the same quality assessments as other sources of real-world data.

Proponents of real-world data point to its potential in achieving a 'learning health system', which help inform both patients and clinicians to improve how they make decisions during clinical visits [23]. However, a fundamental disadvantage of real world data according to our interviewees is the absence of contextual information around the patient/service user, the provider and local practice. This was mentioned specifically around the ambition to 'research the effectiveness of interventions or practice'. Missing contextual data means that associations observed in real world data models can be subject to endogeneity as the results may be compromised by omitted variables and unobserved factors governing selection into different treatment regimens. Such contextual data can explain why patients, who according to observed variables are identical in socio-

demographic characteristics, co-morbidities, and disease progression, are prescribed different treatment or care packages. These unobserved variables are likely to be instrumental in determining patient outcomes, although are usually minimised or balanced across groups in randomised designs. Such unobserved factors are also a reflection that real-world data, and certainly those collected on a routine basis, are not collected with a particular underlying theme or research question in mind (this was also a recurrent theme in the interviews). This means that it is often the case that constructs in real world data (especially routinely collected data) can be outdated and data collectors can de-prioritise the collection of constructs or information that are of substantial value for research purposes.

Interviewees did not recommend a particular source of data or statistical technique that could be used to minimise this form of bias and it was recognised that such bias could arise in multiple ways. One recommendation was that any real world data project should involve a clinician to discuss the potential factors around treatment options and why and how variations in treatment regimens arise; while this would not eliminate the issue it would help to understand the magnitude of potential unobserved variables. Real-world data projects should not be undertaken in isolation of clinical expertise; doing so risks ignoring potentially important sources of unobserved heterogeneity. The problem of unobserved confounding variables was not limited to clinical real world data projects, but was recognised as a limitation of social care real-world data projects where unobserved factors around the provision of unpaid care and the unmeasured support being provided represented a substantial limitation: *"putting the person at the centre of package of care means it is sometimes difficult to record those additional elements for that package of care within the statistics that are recorded and measurable… other elements of the package not normally delivered uniquely by local authority: the carer, friends, family all of the non-recordable effects in an individual life"*.

Proof of concept: how should we measure the robustness of real world data?

The main defining advantage of real-world data, besides apparent advantages in terms of cost, sample size and representativeness, is its (ostensibly) high external validity [1]. The external validity reflects both the delivery of an intervention (to a group that is representative of the general population), but more crucially in the delivery of the control, which usually involves an alternative treatment regimen (best available alternative) as opposed to a placebo. While there is an expanding literature citing studies and study protocols that have been conducted using real-world data, interviewees (those from clinical backgrounds) emphasised that real-world data was not a replacement for/did not supersede the findings from RCT studies. Real-world data is prone to forms of epidemiological bias unlikely to be replicated in findings from well designed and executed RCT studies [1 3]; however, as several interviewees pointed out, RCT study data can also be subject to bias, and some felt that observational data was subject to greater scrutiny despite its often superior properties in terms of transparency, than RCT data are. Others highlighted that there was a need for undertaking activities and developing methods aimed at bridging the divide between RCT and non-RCT studies; pragmatic trials may be one way; recruitment of patients who have a greater resemblance to 'real patients' may be another.

While there are examples of studies that have considered undertaking RCTs studies using routinely collected real-world data as a basis for sample selection or otherwise informing the design of the study [1], some of our interviewees felt that more should be done to evaluate the properties of real-world data and to understand the magnitude of potential bias within real-world data studies. Specifically, some viewed that a true mark of the robustness of observational data is the ability to replicate the findings of RCT studies in carefully matched real-world data studies. While other interviewees felt that this notion was an underlying motive behind many real-world data validation studies for example [69] [106], for some this avenue of enquiry, and particularly moving beyond construct validation studies, was an important 'proof of concept' for real-world data studies. However, there are also several reasons why results of RCT and observational data may not be concordant, and separating out the influence of bias from the influence of measuring effectiveness vs efficacy and explaining the 'real-worldness' is difficult. A lack of replication of results of real world data can both symbolise the strength or real-world data, as well as its limitation.

Future directions

Two themes emerged around future potential of real-world data. The first of these is around the expanding potential of pragmatic clinical trials (PCTs). Unlike traditional RCTs, PCTs are trials that take place within real-world environments and among representative samples of patients, thereby placing the focus on establishing the effectiveness of interventions, as opposed to their efficacy. Within a PCT, patients are randomised to receive an intervention or control treatment but the focus on mimicking real-world conditions means that, among other factors: (i) the control treatment provided often represents the best viable alternative already in place (as opposed to a placebo as can be the case in some RCTs), (ii) the patients randomised have reflect the normal range of patients in terms of disease severity, comorbidity and demographic characteristics; and (iii) the measures of effectiveness collected as outcomes are valid and easily understood by a range of stakeholders, including clinicians, patients, policy-makers, and health commissioners. Real-world data collected through electronic health records was viewed as the basis for designing and undertaking a greater number of pragmatic trials (PCTs) and a number of real-world sources theoretically provide the means of implementing studies and monitoring outcomes in real-time. Evidence from PCTs is likely to be of substantial interest to NICE in establishing the effectiveness of interventions in real world settings while maintaining randomisation, thereby eliminating or at least substantially reducing the occurrence of channelling bias; the proliferation of real world data sources may facilitate this form of evidence to become increasingly frequent in the future.

In practice the trial has been useful as it utilises routinely collected health records to run trials to observe patient's responses to drugs…pragmatic trials where you randomise patients and everything is closely monitored in the real world unlike a randomised trial. Pragmatic trials are very different as they represent a broad sample of clinicians and patients. There is a wide range of data collection … whole purpose is in keeping it simple and keeping it real.

A second theme that emerged was around new technologies stimulating new forms of real world data to be collected. Methods of collecting patient reported outcomes were thought to be shifting from paper to digital devices (smartphones and tablets): "*we have a lot of interest in technology where people get messages on their mobile phone to fill out symptoms, whether these are severe and so on. Uptake is very good and this type of model can be utilised for trials quite easily… where you have mobile phone technology sending information you don't have lots of paperwork… modern technology can help a lot with that. Also with ipads there is a strong movement to increases use in that.*"

For me a survey could be sending a message to a patient on a regular basis surveying their experiences. The old model of surveys will be disappearing, keeping it simple. For me in that sense is simple data collection can be collection such as using disease registry

Patient and service user views are underrepresented in most real-world data sources

There are substantial disjoints in the translation processes between establishing the efficacy and safety of new medical interventions through experimental studies and the implementation of such technologies into everyday practice [37]. Key in establishing the effectiveness of new technologies is establishing that the voice and views of patients and service users are represented. All interviewees felt that this was a weakness of routinely collected real-world data. However, some respondents did report that there were increasing efforts underway to secure the linkage of data that more broadly reflects patient wellbeing to existing real-world datasets. For example the National Cancer Data Repository has considered information on quality of life [107]. There are also programs of NHS surveys (some described in this report) measuring patient satisfaction if not 'outcomes' per se. Additionally, this situation is likely to change rapidly over the next few years as the data and the ability to link different forms of data are enhanced. As one interviewee reported: *"Eventually we would like to link [our data on cancer] with mental health, cardiovascular disease; we would love to pull social security records to know how cancer can impact on going back to work or not; it would be wonderful to know."*

At the moment, survey data usually provide indications of public views and attitudes, and these can reflect experiences of the services provided (e.g. satisfaction with the NHS has been collected as part of the British Social Attitudes survey), although these data rarely have the depth of information necessary to measure how views and quality of life varies across different patient trajectories and the necessary sample size to study how these vary across different treatments or care packages. None of our interviewees mentioned sources of qualitative information that were collected routinely as part of real-world data projects.

The lack of patient voices in real world data as they stand does serve as an impediment on the ambition of NICE to holistically research the effectiveness of interventions or practice or to establish quality guidance since this type of data does form an important component for these assessments. The lack of data on patient views and experiences is also at odds

with the potential capacity of real-world data to establish the effectiveness of treatments and care packages.

Patient Consent and Awareness

Some of the experts interviewed discussed issues of consent and in particular viewed the ambiguities around where consent has been gained and patient and service user expectations around how data were used. In relation to how consent is obtained through hospital and GP surgeries it seems at times could be problematic. One interviewee shared the awkwardness felt by some GPs' when asking for consent: *"when patients don't give consent, GP's are nervous and it puts them in an imposition to get consent... when there is uncertainty the easiest thing to do is not to do anything."* A further interviewee highlighted the contrast between public acceptance for data on shopping behaviours to be collected and analysed by large supermarkets to personalise shopping experiences, but comparative reluctance for medical researchers to conduct a similar undertaking. Some experts also felt that the publicity surrounding care.data had also (temporarily) disadvantaged the position of real-world data depositors. Concern around consent is echoed in the literature where in 2004, Black and colleagues found that only a fifth of databases had collected any form of consent or included an opt-out clause [18]. Elsewhere, consent has been described as 'gold standard yet problematic' [3]

The current consensus is that while issues around patient consent in the collection of many sources of real-world data are unlikely to breach the constraints of the Data Protection Act, the ambiguity and lack of patient and service user awareness of where data is collected and how it is analysed does restrict the expanded use of real world data. In particular, one interviewee expressed concern that patient data was being sold and managed by commercial organisations, and much of the public would be unaware of this commercial arrangement. Where patient consent was raised as an issue, it was also felt that this would also be reflected in the quality and breadth of the data: *"there are differences in nature of the studies in terms of consent which are reflected in a way of the quality of the data"*.

> *"The public doesn't like the sounds of real-world data but if the information is communicated that it is used to improve care and that is what is happening it will not put people off. But the sound of big databases farming it [patient data] out puts people off"*

Data for epidemiological purposes

Examples of survey data were usually cited when considering the type of real-world data that may be suitable for establishing epidemiological trends. However, for rare or genetic conditions some viewed registry data as holding potential. One expert gave the example of Familial hypercholesterolemia, a genetic disorder, and the way in which registry data could be used as a surveillance system for monitoring trends and identifying potentially high risk patients. Some of the certain markers of the disorder are high levels of

cholesterol and heart failure occurring among younger people aged 45 years and under. Heart failure registry data could be used to identify these cases where there was a high likelihood of the disorder among individuals and procedures around cascade testing from these individuals could be used then to identify other family members likely to be susceptible.

Specific issues in social care data

Real world data in social care

The interviewees who were expert in social care data sources reported that 'real world' data was a less frequently used term within social care circles. In the literature, some have claimed that in the case of social care real world data, the data are limited due to the difficulty in identifying meaningful outcomes for social care service users [108], and that this has led to data collection being focussed on information on outputs and costs more so than the type of data needed to fully assess and understand the effectiveness of different social care models. Social care real world data sources were more likely to include survey data, although even these were perceived as having disadvantages in the breadth of the additional contextual data that was included and that they were: *"able to offer some level of baseline information…but there are restrictions on what information they can give; real world data are good at giving statistical elements but not good giving the things that aren't recorded…more of the social side and we're obviously more interested in this in the social care sector"*. Therefore the multiplicity of care sources, particularly unpaid care sources, that are usually part of an individual's care package are difficult to assess and measure and *"makes it difficult to assess 'real worldness' of the data."*; in addition, this limited the extent to which new and emerging problems in social care delivery could be investigated, such as the increased risk of financial abuse following moves towards personalised budgets.

Aggregate vs individual level data

For social care in particular, our interviewees reported that there are substantial challenges in obtaining individual level data (as opposed to aggregate data) that allow for re-analysis and full secondary data analysis. While the data are initially collected in individual form, they can only be deposited into national repositories in aggregate form (which is viewed as an unnecessary burden on Local Authorities who input the data, and a hindrance to researchers who want to use the data). Large-scale survey data provides complementary information but provides only small sample sizes of health and social care service users. Data sharing arrangements with individual Local Authorities are one way of overcoming the challenge of obtaining service-user level data, although establishing these arrangements are time consuming:

> *"Social care services - 152 Local Authorities! They're happy to set agreements and we can get their data if we help them to understand the advantages - but it's very time consuming and complicated!"*

A patchwork quilt of social care data

Interviewees described that real world social care data was like a patchwork quilt. Unlike clinical data where many broad research questions are addressed using a single source of data, some of our interviewees reported a mosaic of different sources was necessary to address social care research questions. One interviewee described the process of finding social care real world data akin to 'detective work' and would consult with a variety of secondary sources in order to uncover new or alternative sources of real-world data. Another interviewee described that this situation was unlikely to change in the near future as there was a lack of strategic coordination and leadership in coordination issues around social care data in the sector: *"social care is underrepresented in real-world data per se; we have HSCIC but there are differences in between what is collected and how we measure effectiveness ….Historically far more money was put into pharmaceuticals research – and interest in health research than health and social care, and this is reflected in the availability and quality of data."* However, a new strategy around care information that has been developed by HSCIC may go some way in addressing this gap[21].

Private sector sources of data do exist and are used but have limitations that are not always shared with public sector data

Expert interviewees did not express consensus in the availability and representativeness of data. While some claimed that "*you can get just as much private sector information as public you need to know where to look*", others were more sceptical about the representativeness of social care data in terms of private providers. In particular, many of the real-world data sources that were named did not provide, and in many cases collect, individual level data and focussed instead on aggregated data (e.g. residential care home returns may have been available but data needed to track individual level transitions may be unavailable). However, there are cases in the literature where real-world data collected routinely on an individual basis from private providers have been used, for example in comparing health and social care outcomes between domiciliary care recipients and outcomes for residents of housing with care [109]. These alternative sources may also be subject to different data collection protocols and are likely to differ substantially in terms of quality.

Overall, sources of individual level data were identified as being more prone to being unrepresentative of the experiences of privately funded social care service users. For example, while the Adult Social Care Survey collects data on social care users' and carers' experiences from a large-sample, this does not include those who self-fund, who account for up to 45% of those in residential care [110].

> *"The social care system in England is means tested and there are a substantial number of older people who are excluded on the grounds of income and wealth. There are very few data sources, ELSA being one of them, that tell us anything about what is happening to those people"*

[21] See https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/443353/HSCIC-Strategy-2015-2020-FINAL-310315.pdf

One interviewee contrasted the situation in the UK regarding real-world data with that of Norway and Scandinavia where data were obtained through "*Statistics Norway across a group of sectors; you can pay a small fee for individual level data*" and concluded that in the UK "*while the real world data offers potential opportunities for studying real world practice, in reality a lot of this is unobserved in the data.*"

Specific issues in clinical data

Recommendations around disease registry and clinical audit data

When asked how NICE should research and evaluate the effectiveness of interventions, responses tended towards recommending data collected through registries and collected for the purposes of clinical audits as holding untapped potential in considering how the effectiveness of interventions should be measured. These data were identified as holding more detailed accounts of treatment regimens. However, where these data were thought to be particularly useful was when they were linked to clinical database data, such as HES or CPRD, for the longer-term monitoring of serious adverse events. Some also viewed audit data as being well utilised in monitoring the implementation of guidelines around treatment and care, but underutilised in the development of guidelines. Registry and clinical audit data for relatively common conditions such as heart failure or diabetes were viewed by interviewees as useful sources of real world data.

This contrasts with some views expressed in the literature which suggest that disease registers are only useful for rarer conditions. In particular, the depth of measures and the direct remit of audits in improving patient outcomes and clinician performance made audit data particularly useful in the mind of one interviewee.

> *"clinical audit data is special as we measure what is useful to measure, not what is easy to measure"*

Some therefore described that register and audit data was more akin to a real-world data set that was designed with a specific set of research questions in mind; they also described their usefulness in monitoring new or less common treatment options. However, even within registry data, there may be variations in the quality of the data available; for example: *"with lung cancer because we have the audit early on data on lung cancer, it is very good... if you go to some other cancers – at the moment some data is not so good at a population level."*

Registry and audit data are also not immune to the critiques around the difficulties in linking data, and some interviewees viewed that different registries were particularly weak at communicating with one another. Nevertheless some did not view challenges in linking data as being insurmountable and described steps that were being taken to enable the data to be linked: *"If you have a particular question you can pay to link these systems*

and pull out data [GP data]. These things are all in this 'Black Box' of data [which can be linked to specific registry data]. At the moment two systems sit on top of it - one for the analyst and the other is meant for more general user - all in development - not publically available and the quality of data varies. … In time will all be linked once quality assurance is more secure… say in a year or 18 months it will be literally one very large spreadsheet 11 half million or 12 million people in there with 10's of millions of records. [It] will be spectacular; nowhere in the world will there be anything like this."

Disease coding frames and accuracy

Standard disease coding frames are often used to highlight the virtues of real-world data. They can help researchers to distinguish co-morbidities from complications in real-world data and the next iterations may be rolled out with greater ability to examine healthcare related injury (iatrogenic diseases) [36]. However, while some interviewees did see the value of standard disease coding frames, the level of granularity of these was questioned. For example ICD-10 codes (International Classification of Disease codes) are used across several datasets and were perceived as an aid for national comparison studies. However, because they are used in developing and developed world contexts, some interviewees questioned the level of granularity that these codes could achieve in terms of many non-communicable/age-related/chronic diseases and their interventions. Registry data, because of their specialised nature, were said to be less sensitive and collected greater information on the disease which could help to overcome this limitation. Other studies have questioned the granularity of information on treatment options [36].

A reliance on CPRD data?

Interviewees described that GP data was held by a number of different companies and organisations, with their estimates ranging from 6-9% to 15% of their market share. CPRD was described as being one of the first organisations to establish itself as a provider of GP level data for research purposes. Interviewees did mention a number of other sources including – QResearch, Apollo, Systm1, EMIS and others – some of which represent the data collection or extraction software and some of which, e.g. QResearch represent organisations with a similar remit to CPRD and may represent alternative suppliers for NICE. In addition, different providers may have different quality systems – for example the patient lists that GPs hold can have errors: patients frequently change address without notifying their GP and sometimes patients on the list may no longer be present in the area creating variations and duplications, and lead to inaccuracies in establishing denominators and calculating rates. In inner city areas where there is a higher level of population churn this problem can lead to substantial levels of inaccuracies [2]; different providers may have different systems and algorithms in place to deal with these. Similarly, different providers may have different systems in place to deal with selection effects of GP surgeries opting in or out of data capture systems. When asked whether any one provider (e.g. CPRD) held advantages over another, interviewees did not specify this to be the case.

> *"There are differences in how the data are captured and collected – e.g. in terms of free text data rather than coded data….but I can't say one system is better than another."*

Unless specifically linked, all GP level data such as CPRD and QResearch data are likely to have little information on hospital contacts and services provided beyond the limited information contained in discharge letters.

Specific issues in public health data

Difficulties in obtaining data with sufficient granularity at a Local Authority level

One interviewee highlighted difficulties in obtaining data on public health issues at a local authority level. Given that most public health functions are now devolved, difficulties were identified in obtaining data that could help local policy-makers to make informed commissioning decisions that reflected local health trends. From NICE's perspective, this lack of data could hamper its ambitions to measure, for example, the implementation of its guidance, as different Local Authorities can exercise different policies with regards to which services are commissioned and the foci of these services.

Section 4: Selected in-depth data profiles

The English Longitudinal Study of Ageing

Aims and description: The English Longitudinal Study of Ageing (ELSA) is a multidisciplinary study of older people aged 50 and above who are living in private households in England at the time of recruitment. It aims to "measure outcomes across a wide range of domains and to provide high-quality multidisciplinary data that can shed light on the causes and consequences of outcomes of interest"[22]. It collects a wide range of data on socioeconomic position, finances and pensions; social exclusion and participation; health, wellbeing and social care needs and service usage; living arrangements and housing; and retirement processes. In addition to self-reported measures, selected sweeps collected data through a nurse visit which can involve the collection of a number of measures of physical health including lung function tests, grip strength and blood assay measures. While multidisciplinary in nature, ELSA is likely to be of greatest value as a tool for monitoring population health trends among a cohort aged 50 and over and in terms of examining the effectiveness of some clinical, public health and social care interventions.

Background, history and study design: ELSA was developed as a companion study to the US Health and Retirement Study to document the ageing process and the interconnectedness between life domains including health and wellbeing, socioeconomic status, working life and retirement, social and personal relationships, and numerous other domains [111]. Sample participants were originally recruited from the Health Survey for England and data were first collected in 2002/3; since then a further five sweeps of data have been collected from the same participants with a seventh wave collected in 20014/15. The original sample has also been replenished four times (including at the latest sweep) to ensure that the data continue to represent the ageing process among those aged 50/55 years and over. Data from study members' partners has also been collected as part of the study.

Validity of measures (e.g. construct, content, criterion validity) and case definitions: ELSA is a multipurpose study which has included a number of different scales (sometimes only subscales) and validated measures as part of its core set of measures that are fielded at each sweep, as well as validated measures that have been used at irregular intervals. These validated measures cross a number of domains including wellbeing (for example CASP Quality of Life Scale [112]), mental health (for example CES-D measures of depression [113]) and clinical measures, for example the Rose Angina Questionnaire and the MRC Questionnaire. There have been a number of exercises aimed at validating some of the self-reported measures collected; for example validating self-reported levels of physical exercise through accelerometer data [114], while others have used evidence from different longitudinal studies, for example to understand the validity of self-reported coronary heart disease status (CHD) [115]. Validation of measures can also take place within the study

[22] http://www.elsa-project.ac.uk/about

– for example in validating report of self-declared diabetes and diabetes diagnosed through blood glucose tests [116]. Objective measures are also collected through nurse visits and efforts are made to improve the accuracy of the data collected: for example waist circumference is measured twice, and even a third time when the two first measures differ by over 3 cm [117] and similar procedures are also undertaken for other measures such as the use of spirometers that detect whether technically acceptable measures are collected [118]. Some measures may need to be interpreted differently to reflect the way in which they are collected when undertaking comparative work between surveys, although this doesn't compromise validity per se; for example cortisol levels are collected through hair samples provided during nurse visits in ELSA and tend to reflect chronic levels of stress, but are collected through salivary measures in some other studies, including the National Child Development Study (NCDS) and as such need different interpretation.

Clear case definition is likely to be possible using a number of (mainly self-reported) measures that correspond with validated scales, although may be considered more reliable where the data have been linked with others (see below). Clear case definition for some constructs such as social care needs is also possible through the use of measures such as Activities of Daily Living Measures and Instrumental Activities of Daily Living measures, which may correspond with measures across other surveys [119]; while these latter measures will hold less relevance compared to Local Authority measures of social care need, these data nevertheless lend themselves to summarily exploring levels of unmet social care needs and geographic variations.

All case definitions are subject to a similar limitation regarding the length between sweeps of data collection, which at two years is less than some other major sources of longitudinal data, but nevertheless may hold implications for creating population-level epidemiological estimates. This may be especially relevant for unstable constructs expected to change relatively rapidly over time, as well as for domains that can be influenced by changes in policy (such as changes in social care eligibility) or other contextual influencers.

Representative of populations and settings: ELSA is intended to represent residents living in private households in England. New sweeps of data have included new cohorts of participants to ensure that the sample remains broadly representative of people aged 50/55. Sample weights are provided to ensure that analysts can generate estimates that remain broadly representative accounting for attrition and differential response rates to different elements of the study – for example different weights are provided for main self-completion survey, the self-completed sexual behaviour questionnaire, the blood test, the nurse visit and the main survey [118]. In addition, different weights are constructed for longitudinal and cross-sectional analyses. Typically, these weights are constructed to ensure that estimates reflect either the longitudinal sample or broader populations in terms of age, region of residence, household composition, presence of a long-term illness, ethnicity (measured as white vs non-white) and self-reported health.

There are limitations to the study design. One of the main limitations to using ELSA is that it is representative only of those in private households, and the aim of the study does not include representing the experience of residents in community establishments. ELSA did not seek to oversample groups that are traditionally either underrepresented or

underpowered in surveys, for example Black and Minority Ethnic groups or Lesbian, Gay, Bisexual or Transgender people. Where analyses have attempted to examine differences using such groups, they have been accompanied by a number of caveats and have resorted to analysis that overlooks the heterogeneity across diverse groupings (see, for example [120]). There are also some issues with examining heterogeneity in ageing experiences among the oldest old population due to the grouping of those aged 85+ to ensure confidentiality.

Representativeness of different disease stages: This is a multipurpose study and its aim is not to provide detailed case histories of disease trajectories, although some the data do lend themselves to study longitudinal changes and variations in disease/condition status and manifestations. The data have also been used to monitor the quality of care received by study members across a range of indicators (see [121]). Due to the study design (survey of private households), it could be possible to speculate that some experiences of severe illness and cognitive decline may be underrepresented in estimates. Some studies have found indicative evidence of underrepresentation (through increased risk of attrition) of older people with cognitive decline [122].

Clear ethical frameworks: Full consent is obtained from participants and different forms of consent are obtained for different forms of data collection and linkage. Ethical approval for all the ELSA waves was granted from the National Research and Ethics Committee [123]. Funding for the study is obtained from the National Institute of Aging (USA), Department of Health, Department for Work and Pensions, Office for National Statistics, Department for Transport, HM Revenue and Customs, Department for Communities and Local Government.

Dynamic and adaptable: ELSA is a multipurpose study and has responded to the changing needs of users, researchers and policy-makers through collecting an increasing breadth of information including blood assays, self-reported sexual behaviour and sleep patterns as well as responding to policy changes such as the changes to the state pension age and contextual effects such as the recession. ELSA's principle investigators are based at NatCen, UCL (Epidemiology and Population Health) and the Institute for Fiscal Studies and are supported by a number of other academics and an advisory group consisting of policy-makers and academics.

Data Linkages: Permissions have been sought to link the data from ELSA respondents with Her Majesty's Revenue and Customs (HMRC) / Department for Work (DWP) records on National Insurance contributions, benefits, tax records and pension; permissions have also been sought to link with Hospital Episode Statistics data and NHS Central Register (mortality) and cancer registration data [111 124]. ELSA shares a similar design with other studies of ageing in Europe and the US allowing for international comparisons; efforts have also been undertaken to harmonise the data with a Korean study [121].

Collection of data reflective of real-world conditions (scope): ELSA collects a broad range of data that includes older people's usage of social care and clinical services, the perception of the quality of their experiences, and their outcomes. Previous studies have included those with a focus on measuring limited health literacy as a barrier to colorectal cancer screening [125], perceptions of age and other forms of discrimination in accessing health services [126], and geographic and socioeconomic-based inequalities in access to full joint replacement across England [127].

Sensitivity: Not applicable with respect to detecting iatrogenic diseases. ELSA has collected detailed life history data which could facilitate establishing the sequence of certain events and diseases.

Understanding or reporting of selection/sample composition: There exists detailed documentation across a number of reports that outline the sample selection and replenishment procedures; these also give an indication as to potential issues with selection and representativeness see [128]. As with survey data, analysts can encounter problems with item or module non-response and, reflective of being a longitudinal study, analysts can encounter issues with wave non-response or attrition. Weights are constructed that can be implemented to help correct for the effects of some of these but not fully and not for all types of missingness.

Uniformity in data collection procedures: ELSA developed as a 'companion study' to the Health and Retirement Survey in the US and the data from both sources have been harmonised [121]. ELSA collects data through a blend of Computer Assisted Personal Interviewing (CAPI) and self-completed questionnaires, in addition to Nurse Visit data (not collected at every sweep) and telephone interviews in certain cases.

Steps taken to minimise common forms of bias: Methods are employed by analysts to correct statistically for many forms of bias. Modifications to the questionnaire design have been tested to explore the impact of bias induced through self-reported bias such as optimism bias in self-reported health [129]. The collection and harmonisation of repeated measures collected in the same sweep (described earlier) also reduced the risk of bias in the collection of bias. Nevertheless, ELSA data is subject to many of the caveats arising in working with survey data.

Scope: As described earlier, there is a wide scope for analysing data reflecting outcomes and experiences beyond morbidity and mortality, particularly with respect to measures of mental wellbeing and quality of life and patterns of social care need and usage.

Granularity of treatment/disease data: ICD-10 codes are used to classify primary cause of death but not for morbidity. Otherwise, high levels of granularity exist for many clinical and social care measures allowing for investigations to take place across different scales and sub-scales.

Other considerations/information: ELSA is a prospective cohort of older people which gives an indication of patterns of ageing in England. A number of different study designs have been implemented within the study, and the data are widely used in cross-sectional and longitudinal formats. The survey content does change at sweeps suggesting a good degree or responsiveness and new concerns are reflected, for example an end of life telephone survey was fielded at the latest sweep. News and updates can be found on the ELSA webpage (http://www.elsa-project.ac.uk/) while the anonymised individual-level data are available (**free of charge**) for non-profit purposes and subject to agreement with other conditions to registered users of the UK Data Archive (http://www.data-archive.ac.uk/); alongside the data a full inventory of study documentation including technical guides and questionnaires can also be found. An eighth wave of data collection is planned in 2016/17 following the latest sweep in 2014/15.

ELSA in numbers: Core data: Wave 1 (2002/3): 12,100 cases. Wave 2: 9,433 cases. Wave 3: 9,771 cases. Wave 4: 11,050 cases. Wave 5: 10,274 cases. Wave 6 (2012/13): 10,601 cases.

Summary and Utility for NICE: ELSA was originally proposed as a potential useful source of data on social care. In the latest wave, the data included data on a range of social care and health topics including the occurrence, data of occurrence, impact and medication taken for a number of chronic mental and physical health conditions, adverse health events, health service usage, social care usage, social care needs, care and the provision of care, quality of life and mental wellbeing, all of which are domains highly pertinent to the work of NICE.

Desired use by NICE	Potential suitability[23]
Research the effectiveness of interventions or practice in real-world (UK) settings	Determination of population of interest; reliability of measures; range of intrinsic measures and additional prognostic measures all satisfied with ELSA. Sample size and the non-purposive sampling of healthy and unhealthy subjects means that researching potential practice impacts may be restricted to widespread practices only.

Example study and abridged abstract:

A cost-benefit analysis of cataract surgery based on the English Longitudinal Survey of Ageing: This paper uses the English Longitudinal Survey of Ageing to explore the self-reported effect of cataract operations on eye-sight. A non-parametric analysis shows clearly that most cataract patients report improved eye-sight after surgery and a parametric analysis provides further information: it shows that the beneficial effect is larger the worse was self-reported eye-sight preceding surgery so that those with very good or excellent eye-sight do not derive immediate benefit. Nevertheless, the long-run effect is suggested to be beneficial. Calibrating the results to existing studies of the effect of imperfect eye-sight on quality of life, the impact of cataract operations on Quality Adjusted Life Years is found to be similar to that established in previous studies and well above the costs of cataract operations in most circumstances [4].

[23] Potential suitability here is something of a subjective construct in the absence of a clear research question to be addressed. However,

Audit the implementation of guidance.	Determination of changes in practice may be challenging in the absence of linked data; in addition the gap in data collection may be problematic for some analyses of change over time; the data will give an overall picture and may be particularly informative in terms of detecting long-term changes in practice around social care. **Example study and abridged abstract:** ***Limited health literacy is a barrier to colorectal cancer screening in England:*** To determine the association between health literacy and participation in publicly available colorectal cancer (CRC) screening in England using data from the English Longitudinal Study of Ageing (ELSA). 73% of participants had adequate health literacy skills. Screening uptake was 58% among those with adequate and 48% among those with limited health literacy skills. Limited health literacy is a barrier to participation in England's national, publicly available CRC screening programme. Interventions should include appropriate design of information materials, provision of alternative support, and increased one-on-one interaction with health care professionals. [125]
Provide information on resource use and evaluate the potential impact of guidance in changing resource use	Determination of inputs/resource use may be patchy and may differ according to the condition. Some studies possible. **Example study and abridged abstract:** ***Self reported receipt of care consistent with 32 quality indicators: national population survey of adults aged 50 or more in England:*** Objective: To assess the receipt of effective healthcare interventions in England by adults aged 50 or more with serious health conditions. Main outcome measures: Percentage of indicated interventions received by eligible participants for 32 clinical indicators and seven questions on patient centred care, and aggregate scores. Receipt of indicated care varied substantially by condition. Substantially more indicated care was received for general medical (74%, 73% to 76%) than for geriatric conditions (57%, 55% to 58%), and for conditions included in the

	general practice pay for performance contract (75%, 73% to 76%) than excluded from it (58%, 56% to 59%). Conclusions: Shortfalls in receipt of basic recommended care by adults aged 50 or more with common health conditions in England were most noticeable in areas associated with disability and frailty, but few areas were exempt [5].
Provide information on epidemiological trends	Determination of several conditions possible – the data have also been used to determine the levels of undiagnosed conditions in populations.
	Example study and abridged abstract: ***Prevalence of frailty and disability: findings from the English Longitudinal Study of Ageing:*** Objective: to examine the prevalence of frailty and disability in people aged 60 and over and the proportion of those with disabilities who receive help or use assistive devices. Results: the overall weighted prevalence of frailty was 14%. Prevalence rose with increasing age, from 6.5% in those aged 60–69 years to 65% in those aged 90 or over. Among frail individuals, difficulties in performing activities or instrumental activities of daily living were reported by 57 or 64%, respectively, versus 13 or 15%, respectively, among the non-frail individuals. Conclusions: frailty becomes increasingly common in older age groups and is associated with a sizeable burden as regards difficulties with mobility and other everyday activities [130].
Provide information on current practice to inform the development of NICE quality standards	Determination of inputs/resource use may be patchy and may differ according to the condition. Depth of information on the components of practice needed to establish quality standards may be patchy.
	Example study and abridged abstract: ***The impact of primary care supply on quality of care in England:*** The aim of this study is to assess the relationship between primary care supply and the quality of primary care in England. We use data from the English Longitudinal Study of Aging (ELSA) which provides a panel of individuals aged 50 and over living in England. Wave 2 to 4 (2004 to 2009) include data on a number of indicators developed for assessing the quality of care of the consultations. The survey data are linked to Primary Care Trusts (PCT) level data on primary care supply measured by the number of

GPs in the area of residence and the average distance to the general practice. In the pooled analysis across all 35 indicators, our findings suggest that, after controlling for individual demographic characteristics, socioeconomic factors, perceived health, and area level factors, a larger number of GPs in the area has a statistically significant and positive impact on quality of care, and distance from GP practice has a statistically significant and negative impact [131].

Key References:

Bridges, S., D. Hussey, et al. (2015). The dynamics of ageing: The 2012 English Longitudinal Study of Ageing (Wave 6) Techincal Report. London, NatCen.

Steptoe, A., E. Breeze, et al. (2013). "Cohort profile: the English longitudinal study of ageing." International journal of epidemiology 42(6): 1640–1648.

Community Mental Health Service User Survey/Community Mental Health Survey

Aims and description: The Community Mental Health Service User Survey (CMHSUS)[24] is a cross-sectional survey of adult mental health service users carried out by the Picker Institute Europe on behalf of the Care Quality Commission (CQC). It is part of the National Patient Survey programme that was established by the Department of Health in 2002 and transferred to the CQC in 2009 [132]. The aim of the survey is collect information from users of community mental health services – service users whose needs cannot be ordinarily met in primary care settings – and to understand their views on the care they received and whether/how this care needs to be improved. The results of the survey can be used by NHS trusts to understand their performance and benchmark their performance against other trusts; a scoring system has been developed to facilitate comparisons between NHS trusts.

Background, history and study design:

The CMHSUS is part of the broader programme of work now under the auspices of the CQC that collects information from service users across a range of clinical services and areas. The mental health survey represents a simple (non-stratified) random sample of users. The latest sweep of available data collected in 2014 surveyed adults (18+) in receipt of specialist care or treatment for a mental health condition and had been seen by a participating NHS trust within a set window [625]. This includes people who were receiving treatment through Care Programme Approach (CPA) and those who were not. The 2015

[24] Surveys of patient experiences have also been known as Community Mental Health Service Survey and Community Mental Health Survey in previous years.
[25] http://ukdataservice.ac.uk/

survey, currently underway, adopted a similar approach. Notable exclusions from the survey include people who were only seen once for an assessment, current inpatients receiving treatment for mental health conditions, and people receiving treatment for a specific area such as drug and alcohol abuse, learning difficulties and specialist forensic services [6]. While there may be similarities in the design of survey, the content changed substantially in 2014 to reflect changes in policy, service delivery patterns and best practice, and this hinders comparability for studies seeking to implement a repeated cross-sectional design.

NHS trusts are responsible for implementing the survey in their areas and are given detailed instructions on how to administer the survey [133]. In 2014, the data are weighted by age and gender to account for differences in the composition of responses between NHS trusts; the data are not adjusted to account for any potential differences between responders and non-responders [133], and the weights do not necessarily enable inferences to be made for the whole population in receipt of mental health services. Detailed information on the steps taken to correct for other factors that may compromise representativeness particularly when comparing scores at a trust level, are given in the data documentation. The 2014 sweep achieved a response rate of 29% with responses obtained from 13,787 individuals.

The 2014 survey asked respondents to complete a total of 49 questions ranging from: their history of contact with mental health services, their experience and satisfaction with the contact and service they received at their last contact, questions on how their care is organised, planned and reviewed and their experiences of these interactions, questions about medication they receive and their feeling of involvement in directing their care, broader questions about the support and information people received, overall perceptions of being treated with respect and dignity; and information about service users' demographics (age, gender, religion, ethnicity and others). Not all of these fields are available for researchers to reanalyse, and only age and gender are available to examine trends by demographic characteristics.

Validity of measures (e.g. construct, content, criterion validity) and case definitions:

The survey has a high level of face validity and asks about service user satisfaction and perceptions, and does not contain tools or measures that are diagnostic in nature. The 2014 redevelopment was intended to improve the accessibility of the survey for users through using more collaborative language, included changes to the language ensure applicability across different NHS trusts, and addressed gaps in knowledge through including new questions such as on service users' relationships with staff. These changes in language, while improving validity, do hinder comparisons between 2014 data and data collected in earlier sweeps.

Representative of populations and settings:

The data are a simple random sample of service users. NHS trusts are asked to compile a list of all service users accessing services over a set period and to select around 900 users randomly from this list using Excel. The intended sample size is around 850 per trust but a sample of 900 is drawn to account for any missingness due to incorrect address details, service users now living outside the UK or death. The aim is to obtain a response rate of 40

per cent; overall 29 per cent was achieved. The survey is administered as a postal survey, and those with mobility issues, other physical and sensory impairments, and those with literacy issues may experience difficulty in completing/returning the survey.

The weighting strategy is intended to correct for compositional differences between trusts in terms of their age and gender, as younger people and women are said to be more critical in their responses than men and older people [133]. However, the impact of non-response on the representativeness of the data is unclear and is a caveat. There are also exclusion criteria imposed on who should complete the survey (outlined earlier) which further constrict the generalisability of the results.

Representativeness of different disease stages: As the data are drawn from a sample in receipt of mental health services, to whom a postal survey is administered, it is unclear the extent to which the responses represent the full spectrum of mental health issues and their severity of users of community mental health services. However, linkage of mental health cluster codes has been approved[26].

Clear ethical frameworks: Research Ethics Committee (REC) approval has been obtained for the survey and trusts must not deviate from the guidance provided [133] in order to comply with the ethical approval granted. Service users are assumed to opt-in to being contacted as part of the survey unless they explicitly opt out.

Dynamic and adaptable: The survey has collected different data across sweeps and the 2014 sweep represented a substantial overhaul of the questions used in order to adapt to the needs of users and decision-makers and to reflect changes in policy and service provision.

The study does not have a longitudinal design in of itself, which does limit the potential research designs that can be implemented - time-series/repeated cross-sectional study designs can be implemented where there are limited changes to the questionnaire.

Data Linkages:

Some linkages have recently been granted to expand the scope of the survey (including providing information on mental health cluster codes)[27].

Collection of data reflective of real-world conditions (scope): A broad scope of data is collected that reflect service user experiences. These can be used to understand how service delivery varies between NHS trusts. The data only allow for a limited understanding of how these vary by the characteristics of the service user e.g. demographics and clinical histories. Nevertheless, the data do allow for examining how real-world practice may deviate from guidelines and expected standards. For example, all service users who are eligible for a Care Programme Approach (CPA) should have out-of-hours contingency care plans in place and all service users regardless of CPA status should

[26] http://www.nhssurveys.org/Filestore/CAG_9-07%28b%29_2014_Community_Mental_Health_Survey_inclusion_of_mental_health_care_cluster_final_approval.pdf

[27] http://www.nhssurveys.org/Filestore/CAG_9-07%28b%29_2014_Community_Mental_Health_Survey_inclusion_of_mental_health_care_cluster_final_approval.pdf

be aware of out-of-hours emergency contacts. However, the 2014 report showed that a fifth (21%) of CPA designated service users and almost two-fifths (38%) of other service users reported not knowing who to contact in the event of a crisis occurring outside office hours [6].

Sensitivity: N/A: the data do not allow for the identification of outcomes or conditions.

Understanding or reporting of selection/sample composition: Some potential limitations around the sample are outlined earlier in the description provided on representativeness. There is little information available to develop an understanding of how respondent characteristics vary from those of the target sample.

Uniformity in data collection procedures: There may be some synergies between the Community Mental Health Survey data and other patient satisfaction data; however the validity of such comparisons is unclear. Furthermore, as the questions do change between sweeps, there is also a constricted ability to examine change over time through repeated cross-sectional studies.

Steps taken to minimise common forms of bias: The data may be subject to bias in terms of systematic differences between responders and non-responders. Other forms of bias may be applicable, for example social desirability bias or extreme responding. The study collectors do attempt to minimise possible errors occurring in the sample design including ensuring that trusts implement the same procedures and limits when drawing a sample (including restrictions of age/services used etc) as well as in implementing opt-out clauses and providing information to respondents [134].

Scope and flexibility in research design: In the absence of data linkage, these data are mainly suited to cross-sectional and repeated cross-sectional/time-series designs.

Granularity of treatment/disease data: N/A – no information is provided on the nature of service users' mental health condition. Information is however provided as to whether service users are provided with a Care Programme Approach (CPA), which provides greater integrated support to users who need it[28].

Other considerations/access information: Raw data are available free of charge to registered users of the UK Data Archive for re-analysis, although do not contain the full range of data collected in the interests of confidentiality. Detailed reports of findings are also available from the CQC website and other sources.

CMHSUS in numbers: Extent of sweeps: 2002-present; Sample size in 2014: 13,787 service users; 57 NHS trusts

Summary and utility for NICE: These data are potentially very useful for NICE in being able to monitor changes in practice since the issuing of guidance and in helping to understand current practice in order to inform on the development of quality standards. The data have been used for a similar purpose within NICE and NICE collaborating centres in the past in order to help develop guidance around improving patient experiences [7]. However, there are limitations around the breadth of respondent characteristics and

[28] Information on CPA status has not been provided in the data deposited on the UK Data Archive

issues – for example there may be limited scope for exploring how standards vary by the needs and characteristics of the client group. The availability of the data on the UK Data Archive also mean that NICE are able to directly access the data and conduct its own analysis. If this is not already the case, NICE should seek to develop dialogue with CQC in order to better understand the potential around this and other NHS surveys of user experiences. This survey was selected to be profiled in-depth because of the paucity of other datasets identified during the course of this review that had an explicit focus on mental health. Some caveats may exist around the sample composition; the data also appear to be rarely utilised within the peer-reviewed literature.

Desire use by NICE	Potential suitability[29]
Research the effectiveness of interventions or practice in real-world (UK) settings	N/A – the focus of this survey is on service user experiences more than outcomes
Audit the implementation of guidance.	Yes, these data are particularly effective for this purpose and many of the fields included in the survey can be directly measured against NICE guidance. **Example study: *Changes in who people see:*** "NICE guidance (CG136) states that changes in staffing can be disruptive to care and it is therefore important that services maintain continuity of individual therapeutic relationships wherever possible. Where changes are necessary, people should be provided with appropriate and accessible information about what is happening. Just over two fifths of respondents (41%) said that in the last 12 months, the people they see for their care or services had changed. When asked what impact this had on the care they receive, just under half (46%) of this group said that it 'stayed the same'. Equal proportions (27%) said that 'it got better' or that it 'got worse'. However, 37% of respondents on CPA and 50% of respondents not on CPA said that they did not know who was in charge of organising their care whilst this change was taking place [6].
Provide information on resource use and evaluate the potential impact of	The data can potentially provide a snapshot of resource use based on service users' reports; repeat cross-sectional

[29] Potential suitability here is something of a subjective construct in the absence of a clear research question to be addressed. However,

guidance in changing resource use	studies may provide information on the impact of changing guidance on resource use.
Provide information on epidemiological trends	N/A – in its current form the data are not suitable for this purpose
Provide information on current practice to inform the development of NICE quality standards	There are examples from existing guidance where information from past surveys have been used to form guidance (and around minimum standards). **Example**: "The guidance is the first to focus on improving service users' experience of mental health services rather than on increasing the effectiveness or safety of the interventions they should be offered. Recent UK surveys of service users showed inconsistent practice and sometimes wide variation in performance among trusts, which were two of the factors prompting the development of the guidance. The surveys showed that a large proportion of service users were not given an out of hours telephone number to use and that during assessment many service users were not involved in decision making about their care and treatment and did not have their preferences taken into account. Almost half reported that they were not given a copy of their care plan. Many users of community mental health services reported that they were not given adequate information about medication and care coordination, and many said that they were not getting effective treatment from trusted professionals; a large proportion said they did not have a care review meeting or their physical health needs met." [7]

Key References:

CQC. National Summary of the Results for the 2014 Community Mental Health Survey. London: Care Quality Commission, 2014.

Picker Institute Europe. Guidance Manual for the NHS Community Mental Health Service Users Survey 2014 Oxford: Picker Institute Europe, 2014.

Kendall T, Crawford MJ, Taylor C, Whittington C, Rose D. Improving the experience of care for adults using NHS mental health services: summary of NICE guidance. *Bmj* 2012;344.

Clinical Practice Research Datalink (CPRD)

Aims and description: The Clinical Practice Research Datalink (CPRD) refers both to the eponymous organisation which is described as the 'English NHS observational data and interventional research service' [30] as well as the shorthand description for CPRD-Gold, which describes a specific dataset. CPRD is jointly funded by the National Institute for Health Research (NIHR) and the Medicines and Healthcare Products Regulatory Agency (MHRA). CPRD-Gold (referred to as CPRD from hereon in) is the primary care dataset that can be considered as the successor to the General Practice Research Database. The new designation reflects the CPRD's ambition to link greater number of datasets in order to understand the entire journey of clinical care.

Background, history and study design:

The GPRD started in 1987 (under the guise of the Value Added Medical Products (VAMP) dataset) as GP practices were beginning to adopt computer software as a means of storing data about their patients, although some practices will have uploaded records predating 1987 [135]. In 2012, GPRD was rebranded as CPRD and currently, 674 practices are estimated to be contributing to the CPRD database [9], making the number of currently participating practices larger than THIN but smaller than QResearch. CPRD holds records on the medical histories of 11.3 million patients, 4.4 million who are currently registered[31]. As with both other large primary care databases, CPRD holds a wide breadth of data on the primary care interactions of patients including demographic information, clinical events such as symptomology and diagnosis, immunisation, tests and test results, specialist referrals, referrals to secondary care, hospital admissions and major outcomes, circumstances relating to patients' death [9]. Free text information is not usually part of the data made available to researchers [9], although CPRD can make these available in certain cases [136]. The CPRD can also facilitate the recruitment of practices and patients to take part in studies requiring the additional collection of biosample data or to take part in trials [9]. In addition, CPRD can facilitate the collection of patient reported outcome data from specific cohorts of patients. Furthermore, additional data can be collected on a practice level, and studies have used this latter approach to assess the validity of case definition in CPRD [68].

CPRD and THIN collect data from practices that use Vision software and both are said to have comparable data structures, differing to QResearch which uses EMIS [137] [138]. Small differences do exist based on operating systems and there are indications that Vision may enable faster coding of items (except prescribing information) and therefore a greater number of items are likely to be coded and that practices using Vision may have slightly higher achievement rates for QoF indicators [138]; however, the effect of these small differences on actual usability, breadth of the data, and data quality is not clear in the literature and may be marginal. Given the close history between CPRD and THIN, there are substantial overlaps between CPRD and THIN data, and practices can participate in both databases. A 2012 study that examined 781 practices that were participating in either CPRD (GPRD at the time) or THIN found that 327 (41.9%) were common to both datasets; 286 (36.6%) were included in CPRD only and 168 (21.5%) were included in THIN only.

[30] http://www.nihr.ac.uk/about/clinical-practice-research-datalink
[31] Email communication dated 27th March 2015

Validity of measures (e.g. construct, content, criterion validity) and case definitions: As is the case for both of the other two major sources of primary care data, THIN uses Read code system to record medical information and there are approximately 250,000 read codes used to record patient diagnoses, symptoms and the care that patients receive; previously CPRD had used OMIS (Oxford Medical Information Systems) [139]. Read codes have been in use in the NHS since 1985, although have been variously updated and different versions do exist; the successor to read codes are Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT); read codes and SNOMED-CT can be cross-mapped to ICD-10 codes [140]. CPRD uses additional codes (OPCS4 and BNF) and will include ATC drug coding and MedDRA coding for adverse event reporting in future [136].

As with other large GP-level clinical datasets it is possible to undertake validation studies and explore levels of consensus of CPRD data using data from an alternative clinical databases (THIN and QResearch) as well as from other (linked and unlinked) data sources. For studies comparing CPRD and THIN data, there are potential for additional validation studies based around the overlaps and between datasets examining potential discrepancies between these records at a practice level. Systematic reviews have been undertaken to explore levels of diagnostic validity in the GPRD (CPRD's predecessor). Herrett and colleagues [106] examined the results of 212 validation studies (1987-2008) of GPRD diagnostic validity while Khan and colleagues [69] examined 47 studies. Both reviews concluded that diagnostic validity in the GPRD was high and the potential for misclassification was low. The included studies used a variety of methods to assess diagnostic accuracy, although a gold standard was considered to be those studies that assess validity through re-examination of medical records, hospital records or GP questionnaires [69]. While the overall level of validity was considered good, there were some specific disease areas where the potential for misclassification was higher mainly involving acute conditions [69], and while Herrett and colleagues [106] found that the median value for confirmed cases stood at 89%, some studies reported values as low as 24%. Since both reviews were undertaken and after GPRD became known as CPRD, there has been continued interest in the quality of the data, and these studies have confirmed lower levels of diagnostic validity for less stable constructs, aspects of previous medical history, or prevalence acute conditions. For example, previous smoking status is thought to be underrepresented [141], as is current BMI status [142].

Of particular note and likely relevant to all clinical databases and more widely are the conclusions of a second study of validity and (mis)classification by Herret and colleagues which compared the cases of acute myocardial infarction in the CPRD, HES and MINAP. While positive predictive values were high, relying on data from any one data source led to underestimates of between 25-50% of events compared to using data from all three sources [141].

Representative of populations and settings: Overall, CPRD data is thought to be nationally representative and on average contains data from around 4.4 million active participants. There are some geographic disparities in the data; for example, while 1.5% and 9.1% of patient records in the database come from the North East and Wales respectively [9], these areas account for 4.1% and 4.8% of the national population. By virtue of the sample design there are also some populations that are not represented such as prisoners and those receiving private primary care [9]. Low data quality for some items

prevents full assessment of representativeness of the data; ethnicity for example is present for around a quarter of records which is comparable with the other two clinical databases, although analyses of CPRD suggests that despite low levels of completion, the ethnic profile mirrors that of census data, and the quality of the data has increased substantially among more recent registrants [143]. Data quality is also an issue with other items such as BMI [142].

Nevertheless, the small disparities in composition are not thought to compromise the overall representativeness of the sample, and most studies using the data regard CPRD as being broadly representative of national primary care trends and practice. This is further reinforced by studies comparing prevalence rates of major conditions which find consistency in rates [144]. Most studies therefore use CPRD data as being representative of national trends and outcomes in primary care.

Representativeness of different disease stages: CPRD is a primary care clinical database that allows researchers to monitor the treatment of disease in primary care settings. In this respect, it shares many of the limitations of the other two major sources of primary care data in that the possibility of monitoring different stages of disease beyond primary across different care providers is only possible through data linkages based on common identifiers (see below). However, part of the justification for rebranding GPRD to CPRD was specifically because of the enhanced data linkages that were expected to take place within the dataset.

Data are used to monitor disease stages in the literature and the factors predicting transitions to different stages and conditions; for example Fleming and colleagues mapped trends in the progression of cirrhosis in primary care using data from the GPRD; the utility of the data for mapping disease stages is likely to be highly dependent on the research question being asked. As CPRD is not designed for examining a single disease or condition, monitoring disease progression from a patient perspective may be challenging; however, there may be scope for including patient reported outcome data through fielding additional tools for capturing patient reported outcomes [136] , although there are few examples of such studies occurring in the published literature.

Clear ethical frameworks: CPRD has approval from the National Research Ethics Service Committee (NRES) for all purely observational research using anonymised CPRD data. Researchers wishing to undertake studies using CPRD data are required to submit their proposal to an Independent Scientific Advisory Committee to gain approval; those who wish to collect additional data from patients will likely need to seek further Research Ethics Committee approval. Patients of participating practices are given notice that their practice is a contributing member of CPRD, and notice that they can opt-out if they wish, although the procedures and number of patients withdrawing consent are unclear [145].

Dynamic and adaptable: Adaptation/dynamism in the actual collection of data will be a reflection of development either in the extent/depth of read codes available, the extent of data linkages and changes in the Vision platform and extraction algorithms. The potential for additional instruments to be fielded among patients means that data that better meets the data requirements of the research question can be collected purposefully; in particular this means that data on Patient Reported Outcomes could be collected alongside other clinical measures. Studies that have exploited this adaptability

include O'Meara and colleagues' study of the pharmacoepidemiology of statins that implemented additional blood tests among patients [146]. CPRD data have been used for a number of epidemiological study designs (see below).

Data Linkages: CPRD data are notable in the extensive data linkages that are available, potentially offering a broad resource for examining patient transitions. Data linkage projects can be completed on request from researchers, although established linkages exist with registry and audit data as well as other clinical and social databases. CPRD existing data linkages include those with MINAP data, National Cancer Intelligence Network data and HES data. Area level data are also available including Index of Multiple Deprivation data and Townsend deprivation scores [9]. Further data linkages are planned. Several studies have exploited the data linkages based on CPRD data; these include, for example, McDonald and colleagues' study of community acquired infections among older diabetics in primary care (CPRD) and the consequent utilisation of secondary care (HES) [147]. Not all practices that contribute to CPRD also agree to further data linkage. In addition, these linked data also have their limitations – for example linked HES data will not provide information on hospital prescribing or testing.

Collection of data reflective of real-world conditions (scope): A broad scope of data reflecting primary care interactions between patient and GP are collected in CPRD, and the extensive data linkages allow for further analysis of patient journeys across primary into secondary care. There exist several published studies that examine a broad range of research questions around the frequency, quality, treatment of regimes and outcomes of these, and the broad scope of data is considered a strength of CPRD [9]. Of particular interest to NICE may be those studies that offer insights into resource utilisation in and beyond primary care: such studies include calculations of incidence of hospitalisations and referrals among people with diagnosed chronic lymphocytic leukaemia [148], and calculating the incidence of consultations, prescriptions, referrals, and diagnostic testing among people living with fibromyalgia [149]. In the latter study, the 'real-worldness' of the data meant that researchers were able to comment on GPs' ability and willingness to classify cases as fibromyalgia.

Sensitivity: There are a number of studies that utilise data from CPRD to assess iatrogenic disease or complications arising from therapy. The degree of sensitivity in this respect will be highly dependent on the disease/condition itself.

Understanding or reporting of selection/sample composition: Consultation with study depositors is likely to inform on the impact and procedures for surgeries opting in/out of the study. There is little evidence of any difference in the profile of participating surgeries in CPRD compared to those who do not participate (beyond the data on representativeness presented above). No literature was found that described differences in CPRD participating surgeries that consented to data linkage (e.g. with HES data) compared to surgeries that did not.

Uniformity in data collection procedures: No purposeful data collection method per se – dependent on GP entry as part of consultation and patient care. As described above, uniformity in coding systems used makes CPRD data highly comparable with QResearch and THIN data.

Steps taken to minimise common forms of bias: Sample selection bias (as opposed to selection bias across the database) can occur in studies but can be minimised through utilising a number of analytical techniques including implementing more sophisticated matching procedures see, for example, [150]. Confounding by indication is a risk in using CPRD data for effectiveness studies, although use of appropriate analytical techniques likely to offset the risk in part. Studies using CPRD data have also undertaken forms of sensitivity analysis by different treatment regimes or stratified analyses by risk profile to examine whether their conclusions hold in order to examine potential impacts of confounding by indication for example [151]. However, these strategies are unlikely to fully remove the possibility of this form of bias [152]. Misclassification bias may be minimised through standardised data entry and the use of read codes, although validation studies suggest that there exist potential differences and discrepancies in the classification of disease. Modifications to the sample drawn can help offset the potential impact of other forms of potential bias such as protopathic bias see, for example [153]. Selection effects at the practice level (and potentially the patient level) are also possible, although the risk is likely equivalent to other clinical databases. As is common across clinical datasets, the influence of unobserved confounders including over-the-counter medications will be unquantifiable as will levels of compliance with reported treatment regimes.

Scope and flexibility in research design: As described earlier, there is a wide scope for analysing data reflecting outcomes and experiences of morbidity and mortality at primary care level, as well as trends in the care and treatment provided. These data can also be linked to many other data sources allowing for potential tracking of patient journeys between primary and secondary care as well as greater detail of the nature of these experiences. There are also examples of studies that have implemented additional data collection through CPRD in order to address specific research questions.

Granularity of treatment/disease data: Given that read codes form the basis for the entry of data into CPRD, and that over 250,000 codes exist, a high level of granularity of data can be incorporated into studies. This granularity has been exploited in many studies to better distinguish between certain forms of condition.

Other considerations/access information: Training is available from CPRD. Costs are associated with accessing CPRD data and vary by the size of the dataset and the nature of the linkages. However, standard extracts approximate at around £15,000 for a dataset of 1,000 patients to over £60,000 for a dataset of over 300,000 patients. An annual license is also available costing between £125,000-£255,000 depending on the nature of the subscribing organisation. The data are supported by a Knowledge Centre who can address most researchers' queries (kc@cprd.com).

CPRD in numbers: 685 GP Practices; 13.7 million patient records; 4.4 million active patients; 85.8 million patient years of data; 8.9% of the population; 75% of contributing practices in England in CPRD with linked data.

Summary and utility for NICE: CPRD data have utility for NICE through the flexibility in being able to collect additional fields and the potential to conduct research based on free-text fields. CPRD data are also available to medical researchers based outside UK universities potentially expanding the pool of potential partners with which NICE could

work in using the dataset. The long established nature of CPRD (based on GPRD) means that several retrospective studies could also be potentially conducted using these data.

Desire use by NICE	Potential suitability[32]
Research the effectiveness of interventions or practice in real-world (UK) settings	Determination of population of interest and reliability of measures are satisfied with CPRD; range of intrinsic measures and additional prognostic measures (although data quality may be problematic for some measures in common with other similar sources of primary care data). Additional scope for collecting patient reported outcomes is possible.

Example study and abridged abstract:

"***Effectiveness of maternal pertussis vaccination in England: an observational study***: Background

In October, 2012, a pertussis vaccination programme for pregnant women was introduced in response to an outbreak across England. We aimed to assess the vaccine effectiveness and the overall effect of the vaccine programme in preventing pertussis in infants. We undertook an analysis of laboratory-confirmed cases and hospital admissions for pertussis in infants between Jan 1, 2008, and Sept 30, 2013, using data submitted to Public Health England as part of its enhanced surveillance of pertussis in England, to investigate the effect of the vaccination programme. We calculated vaccine effectiveness by comparing vaccination status for mothers in confirmed cases with estimates of vaccine coverage for the national population of pregnant women, based on data from the Clinical Practice Research Datalink.

Findings: The monthly total of confirmed cases peaked in October, 2012 (1565 cases), and subsequently fell across all age groups. 26 684 women included in the Clinical Practice Research Datalink had a livebirth between Oct 1, 2012 and Sept 3, 2013; the average vaccine coverage before delivery based on this cohort was 64%. Vaccine effectiveness based on 82 confirmed cases in infants born from Oct 1, 2012, and younger than 3 months at onset was 91% (95% CI 84 to 95). Vaccine effectiveness was 90% (95% CI 82 to 95) when the analysis was restricted to cases in children younger than 2 months. Interpretation: Our assessment of the

[32] Potential suitability here is something of a subjective construct in the absence of a clear research question to be addressed. However,

	programme of pertussis vaccination in pregnancy in England is consistent with high vaccine effectiveness. This effectiveness probably results from protection of infants by both passive antibodies and reduced maternal exposure, and will provide valuable information to international policy makers [154].
Audit the implementation of guidance.	Determination of changes in primary care practice possible using CPRD data; such studies have included studies based on the introduction of NICE guidance.

Example study and abridged abstract:

"***Comparison of cancer diagnostic intervals before and after implementation of NICE guidelines: analysis of data from the UK General Practice Research Database***: Background: The primary aim was to use routine data to compare cancer diagnostic intervals before and after implementation of the 2005 NICE Referral Guidelines for Suspected Cancer. The secondary aim was to compare change in diagnostic intervals across different categories of presenting symptoms. Methods: Using data from the General Practice Research Database, we analysed patients with one of 15 cancers diagnosed in either 2001–2002 or 2007–2008. Putative symptom lists for each cancer were classified into whether or not they qualified for urgent referral under NICE guidelines. Diagnostic interval (duration from first presented symptom to date of diagnosis in primary care records) was compared between the two cohorts. Results: Patients who presented with NICE-qualifying symptoms had shorter diagnostic intervals than those who did not (all cancers in both cohorts). For the 2007–2008 cohort, the cancers with the shortest median diagnostic intervals were breast (26 days) and testicular (44 days); the highest were myeloma (156 days) and lung (112 days). The values for the 90th centiles of the distributions remain very high for some cancers. Tests of interaction provided little evidence of differences in change in mean diagnostic intervals between those who did and did not present with symptoms specifically cited in the NICE Guideline as requiring urgent referral. Conclusion: We suggest that the implementation of the 2005 NICE Guidelines may have contributed to this reduction in diagnostic intervals between 2001–2002 and 2007–2008. There remains considerable scope to achieve more timely

	cancer diagnosis, with the ultimate aim of improving cancer outcomes [8].
Provide information on resource use and evaluate the potential impact of guidance in changing resource use	Determination of inputs/resource use is possible at primary care level as well as linkages into secondary care. **Example study and abridged abstract:** "***The prevalence and incidence, resource use and financial costs of treating people with attention deficit/hyperactivity disorder (ADHD) in the United Kingdom (1998 to 2010)***: The aims of this study were to characterise the epidemiology of diagnosed ADHD in the UK and determine the resource use and financial costs of care. Methods For this retrospective, observational cohort study, patients newly diagnosed with ADHD between 1998 and 2010 were identified from the UK Clinical Practice Research Datalink (CPRD) and matched to a randomly drawn control group without a diagnosis of ADHD. The prevalence and incidence of diagnosed ADHD were calculated. Resource utilisation and corresponding financial costs post-diagnosis were estimated for general practice contacts, investigations, prescriptions, outpatient appointments, and inpatient admissions. Mean annual total healthcare costs were higher for ADHD cases than controls (e.g. £1,327 versus £328 for year 1). Conclusions: The prevalence of diagnosed ADHD in routine practice in the UK was notably lower than in previous reports, and both prevalence and incidence of diagnosed ADHD in primary care have fallen since 2007. Financial costs were more than four times higher in those with ADHD than in those without ADHD[155].
Provide information on epidemiological trends	Determination of several conditions possible. Large population samples of patients with rare conditions are possible using these data (although not in all cases). **Example study and abridged abstract:** "***The incidence of pneumonia using data from a computerized general practice database***: Despite being widely recognized as a significant public health problem

there are surprisingly few contemporary data available on the incidence of pneumonia in the UK. We conducted a general population-based cohort study to determine the incidence of pneumonia in general practice in the United Kingdom. Data were obtained from The Health Improvement Network (THIN) – a computerized, longitudinal, general practice database. Recorded diagnoses of pneumonia between 1991 and 2003 were used to calculate the incidence of pneumonia stratified by year, sex, age group and deprivation score. The overall incidence of pneumonia was 233/100 000 person-years [95% confidence interval (CI) 231–235] and this rate was stable between 1991 and 2003. In conclusion, pneumonia is an important public health problem and the incidence of pneumonia is higher in people at the extremes of age, men and people living in socially deprived areas." [156]

Provide information on current practice to inform the development of NICE quality standards

Determination of inputs/resource use and components of practice likely to be sufficient.

Example study and abridged abstract:

Regional and temporal variation in the treatment of rheumatoid arthritis across the UK: a descriptive register-based cohort study: Objectives To describe current disease-modifying antirheumatic drugs (DMARDs) prescription in rheumatoid arthritis (RA) with reference to best practice and to identify temporal and regional trends in the UK. Participants with RA were identified through screening of all patients in the General Practice Research Database (GPRD) with a clinical or referral record for RA and at least 1 day of follow-up.

Results Of the 35 911 patients in the full RA cohort, 15 259 patients (42%) had incident RA. Analysis of prescribing in incident RA patients demonstrated that between 1995 (baseline) and 2010 there was a substantial increase in DMARD, and specifically methotrexate, prescribing across all regions with a less marked increase in combination DMARD prescribing. Conclusions There has been a substantial increase in prescribing of DMARDs for RA since 1995; however, regional variation persists across the UK with relative undertreatment, according to established best practice. Improved implementation of evidence-based best clinical practice to facilitate removal of treatment

variation is warranted. This may occur as a result of the implementation of published national guidance[157].

Key References:

Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International journal of epidemiology* 2015:dyv098.

QResearch

Aims and description: QResearch is a collaboration between Nottingham University and EMIS (software developers) which provides access to anonymised primary care patient-level data to academic researchers based in UK universities. The aims of this collaboration are stated on the QResearch website as being 'to develop and maintain a high quality database of general practice derived data for use in ethical medical research'[33]. A second aim is to make health morbidity statistics available to the health community at large (at aggregate level) [158]. QResearch is profiled here along with THIN and CPRD as one of the three major clinical databases in the UK.

Background, history and study design: The data collected in QResearch are based on patient interactions with their GP and include information on patients' demographics, lifestyle and physical characteristics (height, weight, smoking status), symptoms, clinical diagnoses, consultations, referrals, prescribed medication and results of investigations; data linkages also offer a broader set of potential data for analyses [159]. Only coded data are available from QResearch - excluding free text data and attachments – which decreases the possibility of breaches of confidentiality of patients [66].

EMIS (Egton Medical Information Systems) is one of the largest providers of integrated software for clinicians in the UK and was designed in the 1980s as software written by doctors for doctors with the ultimate aim of improving patient care[34]. EMIS provides software for GPs and other clinicians to store and interrogate patient-level records. The facility to use anonymised patient records for academic research became functional over 2004-2006 although patient records date back much further to the 1990s when the practices involved in QResearch began adopting EMIS systems [158 160]. Practices benefit in two ways from taking part in QResearch: firstly through the wider findings of research based on QResearch and secondly through feedback on their data quality[35]. The latest publications using QResearch data suggest that over 700 GP practices now contribute data [161], although this number may now stand at 950 practices [11] and the data are thought to represent the medical histories of over thirteen million patients [11 137]. QResearch is alone in collecting data from GP practices using EMIS; CPRD and THIN collect data from practices that use Vision software [137 138]. Small differences do exist based on operating systems and there are indications that Vision may enable faster coding of items (except prescribing

[33] http://www.qresearch.org/SitePages/What%20Is%20QResearch.aspx
[34] http://www.emis-online.com/company-profile
[35] Not all practices who use EMIS are included in QResearch

information) and therefore a greater number of items are likely to be coded and that practices using Vision may have slightly higher achievement rates for QoF indicators [138]; however, the effect of these small differences on actual usability, breadth of the data, and data quality is not clear in the literature and may be marginal.

Validity of measures (e.g. construct, content, criterion validity) and case definitions: ***QResearch uses standard codes:*** As is the case for both of the other two major sources of primary care data, QResearch uses Read code system to record medical information and there are approximately 250,000 read codes used to record patient diagnoses, symptoms and the care that patients receive [139]. Read codes have been in use in the NHS since 1985, although have been variously updated and different versions do exist; the successor to read codes are Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT); read codes and SNOMED-CT can be cross-mapped to ICD-10 codes [140].

As is the case with other large GP-level clinical datasets it is possible to undertake validation studies and explore levels of consensus of QResearch data using data from an alternative clinical databases (CPRD and THIN) as well as from other (linked and unlinked) data sources. One such study examined the impact of statins in reducing mortality from ischaemic heart disease, finding 'remarkable consistency' between CPRD and QResearch data in terms of statistical analyses [138]. Studies on other health concerns, including cancer, also suggest a high degree of concurrence between CPRD and QResearch data [35]. Studies on smoking prevalence also suggest that the data closely follow trends in the Health Survey for England [162]; although overall QResearch data has not been validated as extensively as CPRD data. Furthermore, the data may share some of the limitations found in CPRD and THIN based around the nature in which data are collected and the stability of constructs. There is also potential to undertake validation studies against linked data (see later section), and studies that have examined the levels of valid NHS number through which to undertake such linkages have revealed high levels of completeness to facilitate this work [163]. Where a number of QResearch validation studies have focussed is on validating risk prediction scores that are based on simulation models using QResearch data against actual observed events for example [164].

Representative of populations and settings: QResearch is broadly representative of GP practices in the UK. Three forms of selection effect potentially compromise the representativeness of the data. Firstly, although EMIS is the largest provider of information systems to GP practices not all GP surgeries are represented. Secondly, not all GP surgeries who use EMIS are included in QResearch data. Thirdly, not all patients belonging to included surgeries will consent to having their data included: included practices are asked to display notices in their surgeries to inform patients about their participation in QResearch, and patients who do not wish their anonymised data to be included are able to opt out [158]. These potential limitations are likely to be shared across all three major primary care databases. Some (minor) demographic differences may exist between the three large clinical databases. For example, some differences have been uncovered between samples of THIN and QResearch patients, with QResearch patients having a more even representative socioeconomic profile according to patients' postcodes than THIN [165]. Low data quality for some items prevents full assessment of representativeness of the data; ethnicity for example is present for around a third of records which is comparable with the other two clinical databases and analyses of CPRD

suggests that despite low levels of completion, the ethnic profile mirrors that of census data, and this may also be the case for QResearch data [143]. The largest differences are likely to lie in the size of the clinical databases, with THIN being the smallest and QResearch the largest.

Representativeness of different disease stages: QResearch is a primary care clinical database that allows researchers to monitor the treatment of disease in primary care settings. In this respect, it shares many of the limitations of the other two major sources of primary care data in that the possibility of monitoring different stages of disease beyond primary across different care providers is only possible through data linkages based on common identifiers (see below). As discussed earlier, QResearch has high data quality for NHS number, a common identifier across several databases. As QResearch is not designed for examining a single disease or condition, monitoring disease progression from a patient perspective may be challenging and there are no studies that describe QResearch data being used to examine patient reported outcomes.

Clear ethical frameworks: QResearch has full ongoing approval from the Trent Multicentre Research Ethics Committee and research studies which utilise QResearch data need to obtain ethical approval from this committee. Patients of participating practices are given notice of the facility of opting out of the study.

Dynamic and adaptable: There are numerous examples of QResearch data being used for innovative studies; there are several opportunities for improved patient care being exploited through the development and use of risk scores in primary care which have been found to outperform traditional frameworks. Adaptation/dynamism in the actual collection of data will be a reflection of development either in the extent/depth of read codes available, the extent of data linkages and changes in the EMIS platform and extraction algorithms from EMIS for QResearch.

Data Linkages: QResearch data are routinely linked to socioeconomic data (based on patients' postcodes) and cause of death data (based on Office for National Statistics death certificate) [66]. Further data linkage projects have enabled data linkage between QResearch and cancer registry data [163]. There is also work underway to link QResearch data with hospital episode data (inpatient data, outpatient data, maternity data, critical care data) [66].

Collection of data reflective of real-world conditions (scope): A broad scope of data reflecting primary care interactions between patient and GP are collected in QResearch and studies have included: epidemiological studies around the incidence and correlates of smoking cessation [162]; studies investigating potential iatrogenic complications following prescriptions of statins [166]; and studies examining resource use through trends in consultation rates in GP practice [167].

Sensitivity: The data have been used to examine potential iatrogenic complications following statin prescriptions (see above). Triangulation of findings between different data sources may be necessary to confirm findings. The data may allow for and establishing states before administering treatment.

Understanding or reporting of selection/sample composition: Further contact with study depositors is likely to inform on the impact and procedures for surgeries opting in/out of the study. The impact is likely to be similar across all three primary care databases.

Uniformity in data collection procedures: No purposeful data collection method per se – dependent on GP entry as part of consultation and patient care. As described above, uniformity in coding systems used makes QResearch data highly comparable with THIN and CPRD data.

Steps taken to minimise common forms of bias: Confounding by indication a risk in using QResearch data for effectiveness studies, although use of appropriate analytical technique likely to offset the risk in part, and studies using QResearch data have explored the potential extent and impact of indication bias using different methods see, for example, [166]. Modifications to the sample and to the range of data included can reduce the impact of other forms of potential bias, for example protopathic and recall bias see, for example, [35]. Selection effects at patient and practice level are also possible, although the risk is likely equivalent to other primary care databases.

Scope and flexibility in research design: As described earlier, there is a wide scope for analysing data reflecting outcomes and experiences of morbidity and mortality at primary care level, as well as trends in the care and treatment provided. The study depositors state that QResearch data are suitable for case control studies designed to examine risk factors for onset of disease, cross sectional surveys, cohort studies and sample size calculations (for non-observational studies) [11].

Granularity of treatment/disease data: Given that read codes form the basis for the entry of data into QResearch, and that over 250,000 codes exist, a high level of granularity of data can be incorporated into studies. The data depositors/gatekeepers can also provide advice on the level of granularity

Other considerations/information: QResearch is one of the newest sources of primary care information but is rapidly emerging as the largest source with new practices becoming involved and EMIS being the most common form of information management system in use. Further data linkage projects are likely to enhance the potential for tracking patient journeys in future. However, the data have a number of stipulations: firstly while QResearch is being run as a not-for-profit collaboration, some costs are associated with extracting the data (which are dependent on the study itself). Secondly, to access QResearch, a stringent ethics process needs to be undertaken to ensure that the purposes for which the data are to be used adhere to the underlying principles of QResearch. Finally, and most fundamentally for NICE, data for QResearch are available only to researchers based in UK universities, which compromises its potential for internal use.

QResearch in numbers: 950 GP Practices; 13 million patient records; Largest sample available: 100,000.

Summary and Utility for NICE: QResearch is of interest to NICE for many of the real world data uses identified by NICE, but it appears the data cannot be directly accessed by NICE.

Nevertheless, given the substantial potential of these data, NICE should consider ways of developing research projects based on QResearch data in partnership with universities.

Desire use by NICE	Potential suitability[36]
Research the effectiveness of interventions or practice in real-world (UK) settings	Determination of population of interest and reliability of measures are satisfied with QResearch; satisfactory range of intrinsic measures and additional prognostic measures (although data quality may be problematic for some measures in common with other primary care data). **Example study and abridged abstract:** "***Unintended effects of statins in men and women in England and Wales: population based cohort study using the QResearch database***: Objective: To quantify the unintended effects of statins according to type, dose, and duration of use. Design: Prospective open cohort study using routinely collected data. Setting: 368 general practices in England and Wales supplying data to the QResearch database. Results: Individual statins were not significantly associated with risk of Parkinson's disease, rheumatoid arthritis, venous thromboembolism, dementia, osteoporotic fracture, gastric cancer, colon cancer, lung cancer, melanoma, renal cancer, breast cancer, or prostate cancer. Statin use was associated with decreased risks of oesophageal cancer but increased risks of moderate or serious liver dysfunction, acute renal failure, moderate or serious myopathy, and cataract. Adverse effects were similar across statin types for each outcome except liver dysfunction where risks were highest for fluvastatin. A dose-response effect was apparent for acute renal failure and liver dysfunction. All increased risks persisted during treatment and were highest in the first year. Conclusions Claims of unintended benefits of statins, except for oesophageal cancer, remain unsubstantiated, although potential adverse effects at population level were confirmed and quantified.
Audit the implementation of guidance.	Determination of changes in primary care practice possible using QResearch data. An example of a study examining current and changes in diagnostic practice (not guidelines per se) is given below.

[36] Potential suitability here is something of a subjective construct in the absence of a clear research question to be addressed. However,

Example study and abridged abstract:

"***Incidence, prevalence, and trends of general practitioner-recorded diagnosis of peanut allergy in England, 2001 to 2005***: Previous descriptions of the epidemiology of peanut allergy have mainly been derived from small cross-sectional studies. Objective: To interrogate a large national research database to provide estimates for the incidence, prevalence, and trends of general practitioner (GP)–recorded diagnosis of peanut allergy in the English population. Methods: Version 10 of the QRESEARCH database was used with data from 2,958,366 patients who were registered with 422 United Kingdom general practices in the years 2001 to 2005. The primary outcome was a recording of clinician-diagnosed peanut allergy. Results: The age-sex standardized incidence rate of peanut allergy in 2005 was 0.08 per 1000 person-years (95% CI, 0.07-0.08), and the prevalence rate was 0.51 per 1000 patients (95% CI, 0.49-0.54). A significant inverse relationship between prevalence and socioeconomic status was found. Conclusion: These data on GP-recorded diagnosis of peanut allergy from a large general practice database suggest a much lower prevalence in peanut allergy than has hitherto been found. This difference may in part be explained by underrecording of peanut allergy in general practice. Further research is needed to assess the true frequency of peanut allergy in the population and whether there has been a true increase in recent years[10].

Provide information on resource use and evaluate the potential impact of guidance in changing resource use

Determination of inputs/resource use is possible at primary care level. Studies on prescribing trends have been undertaken as well as studies examining other forms of resource use.

Example study and abridged abstract:

Trends in Consultation Rates in General Practice 1995 to 2008: Analysis of the QResearch® database: This report presents the largest longitudinal study of trends in consultations undertaken in primary care and is part of an ongoing series of analyses using the QResearch® database

(http://www.qresearch.org)[168].

Provide information on epidemiological trends	Determination of several conditions possible. Large population samples of patients with rare conditions are possible using these data.

Example study and abridged abstract:

Trends in the epidemiology of chronic obstructive pulmonary disease in England: a national study of 51 804 patients: Aim: To investigate the epidemiology of physician-diagnosed COPD in general practice. Cross-sectional study of 422 general practices in England contributing to the QRESEARCH database. Data were extracted on 2.8 million patients, including age, sex, socioeconomic status, and geographical area. Trends over time for recorded physician diagnosis of COPD were analysed (2001–2005). There was little change over time in the incidence rate of COPD (2005: 2.0 per 1000 patient-years, 95% confidence interval [CI] = 2.0 to 2.1), but a significant increase in lifetime prevalence rate (2001: 13.5 per 1000 patients [95% CI = 13.4 to 13.7]; 2005: 16.8 [95% CI = 16.7 to 17.0]; P<0.001). Conclusion Given the peak in the incidence rate of COPD, we may be approaching the summit of COPD incidence and prevalence in England. However, the number of people affected remains high and poses a major challenge for health services, particularly those in the north east of the country and in the most deprived communities in England. [169]

Provide information on current practice to inform the development of NICE quality standards	Determination of inputs/resource use and components of practice likely to be sufficient.

Key References:

Hippisley-Cox, J., D. Stables, et al. (2004). "QRESEARCH: a new general practice database for research." Journal of Innovation in Health Informatics 12(1): 49-50

Hippisley-Cox, J. (2014). QResearch. Primary Health Care Specialist Group Conference. Stratford-upon-Avon, Warwickshire.

QResearch. (2012). "What is QResearch?" from http://www.qresearch.org/SitePages/What%20Is%20QResearch.aspx.

The Health Improvement Network (THIN)

Aims and description: The Health Improvement Network (THIN) is a collection of anonymised electronic health records collected unobtrusively that are based on GP-patient interactions from over 550 GP practices. THIN is a collaboration between 'In Practice Systems (INPS)' who developed Vision software used by general practitioners (GPs) in the UK to manage patient data (the underlying software used to populate both CPRD and THIN), and IMS Health who then provide access to the data for use in medical research[37]. IMS is part of a global company which is described as serving 'the world's leading pharmaceutical companies and medical research organisations, providing electronic pseudonymised primary care patient data', and were the company leading on the development of the GPRD (the predecessor of CPRD) [170]. As is common across the primary care clinical databases, the ultimate aim of THIN is to improve patient care[38].

Background, history and study design:

Prospective data collection for THIN began in 2002 although data are available for some practices dating as far back as 1987 [171]. THIN data currently account for around 5.7% of the UK population and comprise one of the large medical datasets alongside QResearch and CPRD. Furthermore, although the records of over 3.5 million *current* patients are held in THIN, these represent around 35 per cent of all records held, the remainder being patients who have withdrawn consent, moved practice or died [172]. Patient characteristics including registration details and the patients' age and gender; medical history around diagnoses, treatments and referrals including information for pharamcoepidemiological studies; background and lifestyle related health characteristics including vaccination status, physical characteristics (height and weight) and smoking status; laboratory results; and detailed information around the patient-physician interactions, are collected in THIN; these are held in four distinct databases (practice, patient, therapy and clinical datasets) [173]. Additionally, unlike QResearch, THIN data do include free-text based information [174]; estimates from 2015 suggest that 35 per cent of all comments in medical records have been coded or anonymised and can potentially be used in research. Additional information may also be available including anonymised questionnaires completed by the patient or GP; copies of patient-based correspondence; a specified intervention (e.g. a laboratory test to confirm diagnosis); and death certificates [174].

CPRD and THIN collect data from practices that use Vision software and both are said to have comparable data structures, differing to QResearch which uses EMIS [137] [138]. Small differences do exist based on operating systems and there are indications that Vision may enable faster coding of items (except prescribing information) and therefore a greater number of items are likely to be coded and that practices using Vision may have slightly higher achievement rates for QoF indicators [138]; however, the effect of these small differences on actual usability, breadth of the data, and data quality is not clear in the literature and may be marginal. Given the close history between CPRD and THIN, there are substantial overlaps between CPRD and THIN data, and practices can participate in both databases. A 2012 study that examined 781 practices that were participating in either

[37] https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database
[38] http://www.thin-uk.net/gps/

CPRD (GPRD at the time) or THIN found that 327 (41.9%) were common to both datasets; 286 (36.6%) were included in CPRD only and 168 (21.5%) were included in THIN only.

Validity of measures (e.g. construct, content, criterion validity) and case definitions: As is the case for both of the other two major sources of primary care data, THIN uses Read code system to record medical information and there are approximately 250,000 read codes used to record patient diagnoses, symptoms and the care that patients receive [139]. Read codes have been in use in the NHS since 1985, although have been variously updated and different versions do exist; the successor to read codes are Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT); read codes and SNOMED-CT can be cross-mapped to ICD-10 codes [140].

As with other large GP-level clinical datasets it is possible to undertake validation studies and explore levels of consensus of THIN data using data from an alternative clinical databases (CPRD and QResearch) as well as from other (linked and unlinked) data sources. For studies comparing CPRD and THIN data, there are potential for additional validation studies based around the overlaps and between datasets examining potential discrepancies between these records at a practice level. For example, Lewis and colleagues [171] examined whether well-established epidemiological associations were reproducible in THIN data, finding that they were, with the exception of a known association between aspirin and myocardial infarction (which itself may be a reflection of the 'real-worldness' of the data and the influence of over-the-counter self-medication). They also established that there were no systematic differences between the results of CPRD and THIN participating practices.

Several disease-specific validation studies have been undertaken employing a variety of different methods to establish the validity of these THIN data including examining free-text information, information from additional physician questionnaires, and examining of alternative read codes. Several have reported high levels of validity across a variety of conditions and treatments including Psoriasis, Chronic Kidney Disease, smoking cessation medication prescriptions, Hepatitis C incidence and ischemic cerebrovascular diagnoses [175-179], although with some caveats [177]. Instances of misclassification are more likely to occur with less stable constructs, such as smoking status [180]. No systematic reviews of validation studies of THIN data were identified.

Representative of populations and settings: Some (minor) demographic differences may exist between the three large clinical databases. For example, some differences have been uncovered between samples of THIN and QResearch patients, with QResearch patients having a more even socioeconomic profile based to patients' addresses than THIN [165]; similar conclusions are also drawn in other studies of representativeness which find that 23.5% of THIN patients live in the most affluent quintile [172]. THIN patients are also thought to be slightly older and the data also have minor differences in geographic distribution across UK regions and countries [172]. Nevertheless, these small disparities are not thought to compromise the overall representativeness of the sample, and most studies using the data regard THIN as being broadly representative of national primary care trends and practice.

Low data quality for some items prevents full assessment of representativeness of the data; ethnicity for example is present for around a quarter of records which is comparable

with the other two clinical databases, although analyses of CPRD suggests that despite low levels of completion, the ethnic profile mirrors that of census data [143]. The largest differences are likely to lie in the size of the clinical databases, with THIN being the smallest and QResearch the largest.

Representativeness of different disease stages: THIN is a primary care clinical database that allows researchers to monitor the treatment of disease in primary care settings. In this respect, it shares many of the limitations of the other two major sources of primary care data in that the possibility of monitoring different stages of disease beyond primary across different care providers is only possible through data linkages based on common identifiers (see below). However, the data are used to monitor disease stages in the literature and the factors predicting transitions to different stages and conditions [181]; the utility of the data for doing so are likely to be highly dependent on the research question being asked. As THIN is not designed for examining a single disease or condition, monitoring disease progression from a patient perspective may be challenging; however there may be scope for including patient reported outcome data through fielding additional anonymised questionnaires [174] although there are few examples of such studies occurring in the published literature.

Clear ethical frameworks: THIN has full ongoing approval from the South East Multicentre Research Ethics Committee and research studies. Researchers wishing to undertake studies using THIN data are required to submit their proposal to a Scientific Review Committee to gain approval; those who wish to collect additional data from patients will need to seek further Research Ethics Committee approval[39]. Patients of participating practices are given notice that their practice is a contributing member of THIN [182], although the procedures and number of patients withdrawing consent are unclear.

Dynamic and adaptable: Adaptation/dynamism in the actual collection of data will be a reflection of development either in the extent/depth of read codes available, the extent of data linkages and changes in the Vision platform and extraction algorithms. However, the potential for additional instruments to be fielded among patients means that data that better meets the data requirements of the research question can be collected purposefully; in particular this means that data on Patient Reported Outcomes could be collected alongside other clinical measures. THIN data are suitable for a number of epidemiological study designs included cohort, case-control and case-series[40].

Data Linkages: Recent developments have seen THIN data being linked with Hospital Episodes Statistics (HES) data, providing potential for studying continuity in care between primary and secondary care. A number of patient postcode-based socioeconomic, ethnicity and environmental indicators are available to researchers including Townsend deprivation quintile scores.

Collection of data reflective of real-world conditions (scope): A broad scope of data reflecting primary care interactions between patient and GP are collected in THIN and there are several published studies that examine a broad range of research questions around the frequency, quality, treatment regimes and outcomes of these. These include:

[39] http://www.csdmruk.imshealth.com/our-data/ethics.shtml
[40] http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database/pros-cons

studies that examine the equity of services provided in real-world settings, for example disparities in cancer screening services provided to people with learning difficulties [12]; studies that examine the levels of and factors related to the discontinuation of anti-depressant medication in primary care practice [183]; and studies that examine the impact of policy changes in the diagnosis of Chronic Obstructive Pulmonary Disease [184].

Sensitivity: The data have been used to examine complications of iatrogenic disease in some studies; for example the incidence of cardiovascular complications following iatrogenic Cushing's syndrome [185]. Other studies, have relied mainly on the breadth of data collected on chronic conditions and the sequencing of events and diagnoses to distinguish between complications and co-morbidities for example [186]. The degree of sensitivity in this respect will be highly dependent on the disease/condition itself.

Understanding or reporting of selection/sample composition: Consultation with study depositors is likely to inform on the impact and procedures for surgeries opting in/out of the study. There is little evidence of any difference in the profile of participating surgeries in THIN compared to those who do not participate. Surgeries whose levels of data quality do not meet agreed standards are not allowed to participate in THIN, although the impact of excluding these practices is relatively unknown.

Uniformity in data collection procedures: No purposeful data collection method per se – dependent on GP entry as part of consultation and patient care. As described above, uniformity in coding systems used makes THIN data highly comparable with QResearch and CPRD data.

Steps taken to minimise common forms of bias: Confounding by indication is a risk in using THIN data for effectiveness studies, although use of appropriate analytical techniques likely to offset the risk in part. Studies using THIN data have also undertaken forms of sensitivity analysis and stratified analyses by risk categories to examine whether their conclusions hold, in order to examine potential impacts of confounding by indication [187]. Misclassification bias is thought to have minimal influence on the conclusions drawn from studies using the data [171]. Modifications to the sample drawn can help offset the potential impact of other forms of potential bias such as protopathic bias [188]. Selection effects at the practice level (and potentially the patient level) are also possible, although the risk is likely equivalent to other clinical databases. As is common across clinical datasets, the influence of unobserved confounders including over-the-counter medications will be unquantifiable as will levels of compliance with reported treatment regimes.

Scope and flexibility in research design: As described earlier, there is a wide scope for analysing data reflecting outcomes and experiences of morbidity and mortality at primary care level, as well as trends in the care and treatment provided. These data can also be linked to HES data allowing for potential tracking of patient journeys between primary and secondary care.

Granularity of treatment/disease data: Given that read codes form the basis for the entry of data into THIN, and that over 250,000 codes exist, a high level of granularity of data can be incorporated into studies. This granularity has been exploited in many studies to better distinguish between certain forms of condition – for example in distinguishing between iatrogenic and endogenous forms of Cushing's syndrome see [185].

Other considerations/information: Training is available from IMS. There exists a dedicated research team within UCL – the THIN Database Research Team – who conduct research into cardiovascular disease, mental health, pharmacoepidemiology and other fields of primary care research using THIN data.

THIN in numbers: 587 GP Practices; 12.4 million patient records; 3.7 million active patients; 85.8 million patient years of data.

Summary and Utility for NICE: THIN data hold potential for much of NICE's intended use of real-world data. THIN data have utility for NICE through the flexibility in being able to collect additional fields and the potential to conduct research based on free-text fields. THIN data are also available to medical researchers based outside UK universities potentially expanding the pool of potential partners with which NICE could work in using the dataset.

Key References:

Lewis, J. D., Schinnar, R., Bilker, W. B., Wang, X., & Strom, B. L. (2007). Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiology and drug safety*, *16*(4), 393-401.

Blak, B. T., Thompson, M., Dattani, H., & Bourke, A. (2011). Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. *Informatics in primary care*, *19*(4), 251-255.

UCL THIN Research. The THIN database: UCL, 2015. https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database

Desire use by NICE	Potential suitability[41]
Research the effectiveness of interventions or practice in real-world (UK) settings	Determination of population of interest and reliability of measures are satisfied with THIN; range of intrinsic measures and additional prognostic measures (although data quality may be problematic for some measures in common with other primary care data). Additional scope for collecting patient reported outcomes is possible. **Example study and abridged abstract:** "***Clinical Outcomes and Cost-effectiveness of Continuous Positive Airway Pressure to Manage Obstructive Sleep Apnea in Patients With Type 2 Diabetes in the U.K***: To assess clinical outcomes and cost-effectiveness of using continuous positive airway pressure (CPAP) to manage obstructive sleep apnea (OSA) in

[41] Potential suitability here is something of a subjective construct in the absence of a clear research question to be addressed. However,

	patients with type 2 diabetes (T2D) from the perspective of the U.K.'s National Health Service (NHS). Using a case-control design, 150 CPAP-treated patients with OSA and T2D were randomly selected from The Health Improvement Network (THIN) database (a nationally representative database of patients registered with general practitioners in the U.K.) and matched with 150 OSA and T2D patients from the same database who were not treated with CPAP. The total NHS cost and outcomes of patient management in both groups over 5 years and the cost-effectiveness of CPAP compared with no CPAP treatment were estimated. Initiating treatment with CPAP in OSA patients with T2D leads to significantly lower blood pressure and better controlled diabetes and affords a cost-effective use of NHS resources. These observations have the potential for treatment modification if confirmed in a prospective study" [189]
Audit the implementation of guidance.	Determination of changes in primary care practice possible using THIN data; such studies have included studies based on the introduction of NICE guidance [184].
	Example study and abridged abstract: "***The impact of the 2004 NICE guideline and 2003 General Medical Services contract on COPD in primary care in the UK***: The introduction of the NICE guideline on COPD and the inclusion of COPD in the new Quality and Outcomes Framework (QOF) were designed to improve the care of people with COPD in primary care in the UK. We have investigated whether these initiatives have had an impact on the prevalence of COPD, the recording of spirometry data and the use of combined inhaled corticosteroid/long-acting beta-agonist inhalers. We analysed data from The Health Improvement Network for the year before and after the introduction of the NICE guideline. Following the introduction of the NICE guideline for COPD and the new QOF, there has been an increase in the prevalence of COPD in general practice and a large increase in spirometry data and prescriptions for combination inhalers." [184]
Provide information on resource use and evaluate the potential impact of	Determination of inputs/resource use is possible at primary care level as well as linkages into secondary care.

guidance in changing resource use	**Example study and abridged abstract:** "***Trends in depression and antidepressant prescribing in children and adolescents: a cohort study in The Health Improvement Network (THIN)***: In 2003, the Committee on Safety of Medicines (CSM) advised against treatment with selective serotonin reuptake inhibitors (SSRIs) other than fluoxetine in children, due to a possible increased risk of suicidal behaviour. This study examined the effects of this safety warning on general practitioners' depression diagnosing and prescription behaviour in children. The study identified a cohort of 1,502,753 children (<18 y; registered with GP for >6 m) in The Health Improvement Network (THIN) UK primary care database. Trends in incidence of depression diagnoses, symptoms and antidepressant prescribing were examined 1995–2009, accounting for deprivation, age and gender" [190]
Provide information on epidemiological trends	Determination of several conditions possible. Large population samples of patients with rare conditions are possible using these data (although not in all cases).
	Example study and abridged abstract: "***The incidence of pneumonia using data from a computerized general practice database***: Despite being widely recognized as a significant public health problem there are surprisingly few contemporary data available on the incidence of pneumonia in the UK. We conducted a general population-based cohort study to determine the incidence of pneumonia in general practice in the United Kingdom. Data were obtained from The Health Improvement Network (THIN) – a computerized, longitudinal, general practice database. Recorded diagnoses of pneumonia between 1991 and 2003 were used to calculate the incidence of pneumonia stratified by year, sex, age group and deprivation score. The overall incidence of pneumonia was 233/100 000 person-years [95% confidence interval (CI) 231–235] and this rate was stable between 1991 and 2003. In conclusion, pneumonia is an important public health problem and the incidence of pneumonia is higher in people at the extremes of age, men and people living in socially deprived areas." [156]

Provide information on current practice to inform the development of NICE quality standards	Determination of inputs/resource use and components of practice likely to be sufficient.

Example study and abridged abstract:

Access to Cancer Screening in People with Learning Disabilities in the UK: Cohort Study in the Health Improvement Network, a Primary Care Research Database To assess whether people with learning disability in the UK have poorer access to cancer screening. Four cohort studies comparing people with and without learning disability, within the recommended age ranges for cancer screening in the UK. We used Poisson regression to determine relative incidence rates of cancer screening. Setting: The Health Improvement Network, a UK primary care database with over 450 General practices [12]

Key References:

Lewis, J. D., Schinnar, R., Bilker, W. B., Wang, X., & Strom, B. L. (2007). Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiology and drug safety*, *16*(4), 393-401.

National Minimum Data Set for Social Care (NMDS-SC)

Aims and description: The National Minimum Data Set for Social Care (NMDS-SC) is a database that collects information on the profile of the social care workforce in England. The data collection is organised by Skills for Care and the data hold information on 700,000 care workers (current and former) working in adult social care across 25,000 establishments[42]. Underlying the collection of the data is an ambition to ensure that the adult social care sector has a 'confident, capable and skilled workforce' to ensure the delivery of excellent quality care[43].

Background, history and study design:

Established in 2005, the data are described as the first attempt at implementing a standardised approach to collecting information on employees working across social care establishments [14]. While there has traditionally been a focus on the adult social care workforce, the data also contain information on the workforce providing care in almost 1,600 children's establishments and approximately 1,000 family establishments social

[42] https://www.nmds-sc-online.org.uk/content/About.aspx
[43] https://vimeo.com/121886971

care. In May 2015, data from 24,621 establishments were included in England; these represented a mixture of private establishments (30.1%), voluntary sector (18.6%), Local Authority/statutory controlled establishments (47.3%) and those that did not fit within any particular class (4.0%) [191]. This allows for comparisons to be drawn between the workforce mix across sectors. In addition, data are collected on approximately 220 variables reflecting details around the social care establishment (including size, location, primary client group, sector), the highest qualifications of workers, induction status, job grade and title of workers; gender, disability status, ethnicity and age of workers; contract type and hours; nationality and country of birth; sectoral experience; sickness; pay; and establishment records around recruitment, retention and destination of leavers[192]. These data are uploaded monthly by employers who provide aggregate estimates or provide individual level data for some or all of their employees[193]. Until 2012, participation in the NMDS-SC was optional, although despite being a voluntary levels of participation were high; after this point completion of NMDS-SC has become the 'mandatory workforce data collection tool for the English adult social care sector'[193], although this may not fully apply to all privately registered establishments[194]. Employers submit details of their workforce and in return are able to benchmark their workforce mix according to national trends using a dashboard of indicators. In addition, they are also able to access e-learning modules on pertinent issues such as dementia care and can apply for workforce development funding. Employers also have an option of sharing their data with NHS Choices and the Care Quality Commission[44].

Validity of measures (e.g. construct, content, criterion validity) and case definitions:

Many of the indicators included in NMDS-SC have high levels of construct validity and are not intended to capture complex diagnoses or constructs, therefore validation studies are in the minority. However, there are indications that some of the measures that are included in NMDS-SC do have low data quality – for example Hussein and colleague's study found that around 55% of records in a 2011 extract held valid information on both ethnicity and nationality [195]. Furthermore, explorations of the validity and quality of some of the data are possible through implementing consistency checks[196]. Overall, there remain some differences in data quality for certain domains, including in turnover rates, and data quality can vary systematically by Local Authority[197]. Nevertheless, previous issues around data quality identified in early reviews [198] are thought to be improving over time[34]. The data depositors (Skills for Care) are also making adaptations to improve data quality with regards to measuring the increasing plurality of ways in which social care is planned, purchased and delivered, including improving data capture around non-CQC registered employees and capturing information on the arrangements of direct payment recipients who purchase their social care directly [199].

Representative of populations and settings:

The data are thought to be representative of statutory/Local Authority run care, as all Local Authorities are required to provide data to NMDS-SC. Skills for Care estimate that there are 38,000 Care Quality Commission (CQC) registered providers and employers of adult social care in England and as of December 2014 it was estimated that 55% of these contribute data to NMDS-SC [200]. In total there are thought to be 1.3 million jobs in the

[44] https://vimeo.com/121886971

social care sector [200], and 663,000 workers are represented in the NMDS-SC data [191]. Those purchasing their own care directly – i.e. single person employers who are also the care recipient – are known to be underrepresented in the data[195].

Representativeness of different disease stages: NMDS-SC does not capture information on the care trajectories of social care recipients. It may be possible to trace longitudinal progression in terms of gaining qualifications and professional development in the data.

Clear ethical frameworks: Data are most commonly provided at an aggregate level to maintain the anonymity of individuals and establishments. Anonymised individual level data are available although subject to approval from Skills for Care who 'will make individual judgments on the supply of raw data taking into consideration the nature of the person / organisation requesting the data and the intended use of the data'[45].

Dynamic and adaptable: The NMDS-SC data user group are responsible for considering changes to the core dataset to ensure that necessary fields are captured (see [201] for recent adaptations that the group has considered).

Data Linkages: While data linkages are not specifically mentioned in much of the literature around NMDS-SC, there are examples of studies where NMDS-SC data have been linked to other datasets. Hussein and Manthorpe's [202] study provides an example where NMDS-SC data were linked with Index of Multiple Deprivation data in order to understand antecedent characteristics of patterns of volunteering in social care settings.

Collection of data reflective of real-world conditions (scope): NMDS-SC is a specialist dataset that can provide unique information on the composition, skill mix and development needs of the adult social care workforce. It does not, currently, contain information on the outcomes of social care recipients.

Sensitivity: This is not applicable for NMDS-SC.

Understanding or reporting of selection/sample composition: An outline of the representativeness (provided above); beside underrepresentation of direct payment recipients, there is little evidence that the NMDS-SC data differs systematically from the total care workforce population.

Uniformity in data collection procedures: There is uniformity within the system as all contributing employers are asked to submit their using a standardised portal, although there are irregularities in intervals for data collection [203]. There may be differences as to whether employers provide granular information on some or all of their employees, as opposed to aggregate data[193].

Steps taken to minimise common forms of bias:

Due to the more narrow focus of NMDS-SC, forms of bias discussed elsewhere, such as protopathic bias of confounding by indication, are not relevant considerations in these data. Some bias may occur as employers essentially collect data 'on behalf' of the workforce. This could mean that an employer may make a 'best guess' as to the ethnicity or age of an employee. Further bias may occur in this process as employers may express

[45] https://www.nmds-sc-online.org.uk/content/view.aspx?id=Accessing%20NMDS-SC%20data

digit preference or informed guesses with respect to other fields in the absence of accurate information[204]. The extent to which this occurs has not been studied in-depth in the literature, although is a recognised caveat of using the data.

Appropriate recognition of the hierarchical nature can be found in some studies, for example Hussein's 2010 investigation into pay across the sector [205]; recognition of which can reduce levels of bias in model estimates. Undertaking longitudinal and event history analyses in these data can be difficult due to left and right censoring occurring [203].

Scope and flexibility in research design: Due to the nature of the data, in epidemiological terms these data are better suited to cross-sectional, time-series and cohort studies as opposed to case-control studies.

Granularity of treatment/disease data: The level of granularity of the data is appropriate for the scope of the data. Detailed information is collected on job roles, experience and the characteristics of the establishment.

Other considerations/access information: Raw data are available on request from Skills for Care (Analysis@skillsforcare.org.uk); preference is given to projects that are funded by the Department of Health; conditions of use are attached to these data around use for intended projects and sharing with third parties.

CPRD in numbers: 24,621 social care establishments; 55% of CQC registered providers of adult social care; 663,000 social care workers represented in the data.

Summary and Utility for NICE: NMDS-SC is a specialist dataset suitable for monitoring trends in the social care workforce. This data can potentially help NICE to understand workforce capabilities and undertake preliminary work to understand the feasibility of implementing new standards and guidance in social care settings.

Desire use by NICE	Potential suitability[46]
Research the effectiveness of interventions or practice in real-world (UK) settings	As social care outcomes are not collected in NMDS-SC, it is unlikely that these data are suitable for this purpose.
Audit the implementation of guidance.	The data may be suitable to examine changes following guidance at a workforce level in terms of indicators such as pay, training or necessary skills.
Provide information on resource use and evaluate the potential impact of	Data from NMDS-SC form the basis for the calculation of some unit costs for social care. Such data could form the basis of studies that examine inequities in resource use by

[46] Potential suitability here is something of a subjective construct in the absence of a clear research question to be addressed. However,

guidance in changing resource use	geographic area, for example. Resource use is not linked to individual social care user data. **Example study:** *Unit costs of health and social care*. University of Kent, 2009.[13]
Provide information on epidemiological trends	Determination of epidemiological trends is not possible using these data. However, determination of workforce data to respond to epidemiological challenges is a key strength. **Example study and abridged abstract:** "***The dementia social care workforce in England: Secondary analysis of a national workforce dataset***[14]: Objective: Little is known about the social care workforce supporting people with dementia in England. This article seeks to compare the characteristics of people employed in the social care sector supporting people with dementia with other members of the social care workforce. This article reports on the secondary analysis of a new national workforce dataset from England covering social care employees. Secondary analysis of this dataset was undertaken using 457,031 unique workers' records. There are some important differences between the dementia care workforce and other parts of the social care workforce in respect of the dementia care workforce being more likely to be female, to work part-time, to be employed by agencies and to be less qualified. Many work for medium-sized care businesses and in people's own homes. The findings are set in the context of efforts to increase training and skills. Knowledge of the social care workforce is relevant to care quality and should be borne in mind when planning interventions and commissioning services.
Provide information on current practice to inform the development of NICE quality standards	Determination of potential inputs/resource use likely to be sufficient to form quality standards in terms of staffing although these are unlinked to patient outcomes.

Key References:

Hussein S. Longitudinal Workforce Analysis using Routinely Collected Data: Challenges and Possibilities. London: King's College London, 2012.

Skills for Care. The state of the adult social care sector and workforce in England 2015. Leeds: Skills for Care, 2015.

Health Survey for England (HSE)

Aims and description: The Health Survey for England is an annual cross-sectional survey profiling the physical and mental health status of the English population residing in private residences. As well as headline trends, the breadth of data collected allows the relationships between health status and socioeconomic, sociodemographic and lifestyle factors to be collected. Each annual survey includes a different theme around a particular disease, lifestyle factor or demographic group, as well as including an extensive range of core questions. The survey is carried out by the National Centre for Social Research (NatCen) (with the involvement of UCL, see below) on behalf of the Health and Social Care Information Centre[47].

Background, history and study design:

The first Health Survey for England (HSE) was carried out in 1991 and was conducted by the Office for Population Censuses and Surveys (later the Office for National Statistics). Since then, responsibility for carrying out the survey has transferred to NatCen and the Research Department of Epidemiology and Public Health, UCL [206].

The survey has a broad remit, with a focus that changes annually that can also involve the collection of data through different instruments, and has a complex sampling frame to ensure representativeness (and in some cases sample power) across England. Waves of data collection have focussed on adults aged 16 and over, but since 1995 the study has also included children living in households, although the sampling frame remains focussed on obtaining nationally representative samples of adults aged 16 and over. To obtain nationally representative estimates, a stratified random probability sample design is implemented where postcode sectors are first randomly sampled, followed by addresses within those postcodes. Postcode sectors contain an average of 3,000 addresses and small sectors (<500 addresses) are combined with neighbouring sectors; further stratification is implemented on the basis of socioeconomic status, local authority, and with additional modifications to account for seasonality for further details see, [206]. Where an address is found to contain multiple dwellings, one dwelling is selected at random and where a dwelling has multiple households in occupation, one household is randomly selected [207]. Weights are constructed for analysts using the data that can account for both probabilistic selection and for non-response [207].

[47] http://www.hscic.gov.uk/healthsurveyengland

Core data collected at each sweep since 1994 include general health and longstanding illness; alcohol consumption; smoking; height and weight; and many individual socioeconomic and sociodemographic characteristics [206]. Other data that have been collected with some interruptions include reports of acute disease, accidents, dietary patterns, cardiovascular disease, contraception, use of cycle helmets, lung function, cotinine, blood pressure, and a number of other fields collected periodically as well as data to support particular survey foci. Data are collected by trained interviewers initially and in a second interview by nurses (dependent on the sweep) using computer assisted interviewing; in 2013 a typical interview lasted for 50 minutes for a single person and 60-65 minutes for a couple household and a nurse interview lasted for 30 minutes [208].

Validity of measures (e.g. construct, content, criterion validity) and case definitions:

The HSE includes complex measures, indicators and scales that have been validated through pilot work, are validated against other instruments in the survey, are benchmarked against other data, or whose validity has been established elsewhere. Examples of the internal validation of measures include the Physical Activity and Sedentary Behaviour Assessment Questionnaire (PASBAQ) , which has been validated against accelerometer data collected from a random subsample of patients and was found to have good levels of validity [209]. However there is also evidence that caution needs to be exercised for some data collected; for example parental reports of their children's physical activity levels using HSE instruments are found to have low levels of validity for estimating moderate to vigorous physical activity levels [210]. Such self-reported lifestyle data may be prone to forms of reporting bias (see later sections). Other studies have relied on instruments known to have high levels of validity and reliability; for example Tiffin and colleagues' [211] used the well-established Strength and Difficulties Questionnaire as a measure of mental health to examine the relationship between mental health and obesity.

Additional checks are implemented before the data are released and not all data collected in interviews are considered valid for analysis purposes [212]

Representative of populations and settings:

The data are representative of private households in England. This has limitations in understanding the health of people living in communal establishments (e.g. residential homes) as well as other populations, for example the prison population. In terms of the impact of this on health estimates, given that people living in establishments are more likely to be older and socioeconomically disadvantaged, this could mean that HSE data under represent levels of ill health compared to a population-wide sample [212]. Other studies have found that the (unweighted) data may over-represent women and under-represent men [213]. Nevertheless, studies using HSE data treat the data as being representative of national trends. Study weights and information on sampling units provided with the data enable analysts to account for the design of the survey and for non-response and to produce nationally representative estimates in their studies.

While the data itself may be broadly representative of the population, there may be some difficulty in obtaining sufficient sample size (and statistical power) in order to conduct analysis on smaller groups occurring in the population including, for example, ethnic

minorities. The 2013 sweep of data contained information from 8,795 adults and 2,185 children meaning that analysis by several smaller demographic or geographic groupings are limited and are more likely to be reliant on aggregations of smaller groupings, losing granularity of data. Some previous HSE surveys have been designed to oversample some groups; in the case of ethnicity, the 1999 and the 2004 surveys focussed on the health of ethnic minorities with corresponding boosts to samples of Irish, Black Caribbean, Indian, Pakistani, Bangladeshi and Chinese respondents and these sweeps can be combined to further boost sample sizes [214]. Other sweep-specific foci can, in some cases, also enable similar analyses to be undertaken for other groups. For example, while communal establishments are usually excluded from the data, the 2000 focus on older people collected data from 2,493 residents in care homes [215].

Representativeness of different disease stages: The HSE is an annual cross-sectional survey. It does not allow for longitudinal analysis of individuals across sweeps. However, some forms of longitudinal analyses are possible based on data linkages which could enable the monitoring of progression to different disease stages (see below). Furthermore, the data give cross-sectional information on the prevalence of diseases stages and stages of needs. For example, recent foci on chronic kidney disease allowed for the estimation of different stages of renal disease stage by health authority characteristics [216]. The latest available sweep (2013, at the time of writing) had a focus on social care, and allowed for the examination of different social care needs and provision across the population (Activities of Daily Living and Instrumental Activities of Daily Living) [217].

Clear ethical frameworks: The latest sweep was granted ethical approval from Oxford A Research Ethics Committee [212]. Individual level data are released, although identifiers are removed from the records so that the anonymity of respondents is preserved.

Dynamic and adaptable: The survey is carried out by two partner organisations on behalf of the Health and Social Care Information Centre (HSCIC). The HSIC has consulted on different aspects of the survey, including content and foci in future sweeps which does allow user input into the survey content; the latest sweep highlighted user interest in the inclusion of wellbeing measures [218].

The study does not have a longitudinal design in of itself, which does limit the potential research designs that can be implemented (time-series/repeated cross-sectional study designs can be easily implemented). However, the potential for data linkages does expand on the number of possible study designs [206].

Data Linkages:

HSE data can be linked (for respondents who consent) to the National Health Service Central Register allowing cancer and mortality data to be linked. Respondents are also asked for permission to link HSE data with Hospital Episodes Statistics data; just under four-fifths consented to these linkages in 2009 (78% and 77% respectively) [206]. Additional neighbourhood level data have also been linked based on respondents' residence.

Studies that have used linked data include a study of socioeconomic deprivation and air pollution on the consequent impacts on lung function; this used data linked based on ward residence to conclude that lower social class and poor air quality were independently

associated with decreased lung function [219]. Another example study comes from Oyebode and colleagues [15], who pooled data from sweeps 2001-2008 to obtain a sample of 65,226 records which were linked to mortality records. Cause of death was categorised and the study examined the impact of fruit and vegetable consumption on the risk of death by any cause, death from cancer, and death from cardiovascular events, finding that eating 7+ portions of fruit or vegetables daily reduced the hazard of death by approximately 33 per cent in any given period.

While the HSE raw data are available on the UK Data Archive; special permission is needed to use these linked data which can be granted by NatCen/HSCIC.

Collection of data reflective of real-world conditions (scope): HSE provides comprehensive information on the health of the population in England. This includes collecting a broad set of indicators reflective of real-world conditions in terms of epidemiological conditions, social and socioeconomic determinants of health, health service usage and, since 2011, biannual estimates around social care needs and usage among those aged 65 and over. The data can also be used to investigate how any of these factors vary by social groupings; for example, Nazroo and colleagues' [214] study examined whether inequalities based on ethnicity that occur in the US in health service access and usage were replicated in England; in the main this was not the case although differences did occur in the use of secondary care services. The potential data linkages also make investigating the outcomes of some of these factors and conditions possible. There are also some data collected in HSE that are not collected in other sources profiled in this report; this includes limited information on over-the-counter medications in use by the population; for an example on statins see [220].

Sensitivity: Distinguishing co-morbidities from complications is challenging due to the study design. However, researchers have used the breadth of HSE data to distinguish comorbid conditions as baseline in linked data projects see [221].

Understanding or reporting of selection/sample composition: Some potential limitations around the sample are outlined earlier in the description provided on representativeness. These are outlined in full in the study documentation, along with information as to how analysts can work to mitigate some of these. An additional selection effect could also potentially occur when using linked data as not all respondents consent for their data to be linked to other records.

Uniformity in data collection procedures: Procedures for data collection in HSE mirror those of other large population studies; there is potential for comparison between HSE data and Scottish Health Survey data see [221 for an example].

Steps taken to minimise common forms of bias:

Due to the cross-sectional study design of HSE, forms of bias discussed elsewhere, such as protopathic bias of confounding by indication are, in part, less relevant considerations as the study design would usually preclude carrying out the types of 'effectiveness studies' where these forms of bias would be most pertinent.

Some forms of respondent level bias have been outlined earlier, such as social desirability bias where parents were completing overestimating the amount of exercise undertaken by

their children [210]; this form of bias may also be a factor in adult self-reports of physical exercise [222]. While some of these effects are inevitable in carrying out survey-based research, the extensive detail around methodology and the high number of studies critically investigating and utilising the data does mean that analysts can at least be well informed as to potentially compromising features of the data.

Scope and flexibility in research design: In the absence of data linkage, these data are mainly suited to cross-sectional and repeated cross-sectional/time-series designs.

Granularity of treatment/disease data: Detailed information is collected on variety of domains that allow for distinguishing different diseases and conditions but also allow for a very high level of detail in terms of data on social determinants of health and lifestyle factors; for example in terms of examining the impact of shift work on health status, analysts are able to compare the impact of a total of eight different categories of shift work [223].

Other considerations/access information: Raw data are available free of charge to registered users of the UK Data Archive. Additional permissions and charges are applicable for non-standard or early release data and dependent on the request, these are obtained from NatCen or the HSCIC.

CPRD in numbers: Extent of sweeps: 1991-present; Sample size variations – adults interviewed: 4,645 (2009) – 16,443 (1996); since 2010 the number of adults interviewed has been around 8,500 with around 2,000 interviews carried out with children living in the same household.

Summary and utility for NICE: HSE was suggested in the context of monitoring epidemiological trends. However, the 2013 sweep contained a focus on social care and the survey is a multipurpose survey that can be used to gain an understanding of trends over time in terms of resource utilisation, epidemiological trends, trends in social care needs and usage, trends in lifestyles and social determinants of health, and some trends in prescribing and attitudes to health. The survey data may have great utility for NICE in gathering contextual information critical in the assessing feasibility of different forms of guidance aimed at public health and social care challenges. The data also have the added advantage of being relatively easy to obtain for further secondary data analysis and are free of charge.

Key References:

Boodhna, G., S. Bridges, et al. (2014). (Vol 1): Health, social care and lifestyles. Health Survey for England 2013. R. Craig and J. Mindell. Leeds, Health and Social Care Information Centre.

Boodhna, G., S. Bridges, et al. (2014). (Vol 2): Methods and documentation. Health Survey for England 2013. R. Craig and J. Mindell. Leeds, Health and Social Care Information Centre.

Mindell, J., J. P. Biddulph, et al. (2012). "Cohort profile: the health survey for England." International journal of epidemiology 41(6): 1585-1593.

Desire use by NICE	Potential suitability[48]
Research the effectiveness of interventions or practice in real-world (UK) settings	There is scope to implement repeated cross-sectional designs with (unlinked) HSE data, although this type of study will only give limited insight into the effectiveness of interventions or practice. HSE data that are linked with other sources may offer greater utility and the possibility of implementing a longitudinal design. **Example study and abridged abstract:** "***Fruit and vegetable consumption and all-cause, cancer and CVD mortality: analysis of Health Survey for England data***: Governments worldwide recommend daily consumption of fruit and vegetables. We examine whether this benefits health in the general population of England. Methods: Cox regression was used to estimate HRs and 95% CI for an association between fruit and vegetable consumption and all-cause, cancer and cardiovascular mortality, adjusting for age, sex, social class, education, BMI, alcohol consumption and physical activity, in 65 226 participants aged 35+ years in the 2001–2008 Health Surveys for England, annual surveys of nationally representative random samples of the non-institutionalised population of England linked to mortality data. Results: Fruit and vegetable consumption was associated with decreased all-cause mortality (adjusted HR for 7+ portions 0.67 (95% CI 0.58 to 0.78), reference category <1 portion). Conclusions: A robust inverse association exists between fruit and vegetable consumption and mortality, with benefits seen in up to 7+ portions daily. Further investigations into the effects of different types of fruit and vegetables are warranted.
Audit the implementation of guidance.	There is scope for auditing the implementation of guidance through examining change in practice at a population level; one of the strengths of HSE data in doing so is the ability to examine social or medical inequalities in the implementation of guidance. **Example study and abridged abstract:**

[48] Potential suitability here is something of a subjective construct in the absence of a clear research question to be addressed. However,

	"***Are Current UK National Institute for Health and Clinical Excellence (NICE) Obesity Risk Guidelines Useful? Cross-Sectional Associations with Cardiovascular Disease Risk Factors in a Large, Representative English Population***: The National Institute for Health and Clinical Excellence (NICE) has recently released obesity guidelines for health risk. For the first time in the UK, we estimate the utility of these guidelines by relating them to the established cardiovascular disease (CVD) risk factors. Health Survey for England (HSE) 2006, a population-based cross-sectional study in England was used with a sample size of 7225 men and women aged >= 35 years (age range: 35-97 years). The following CVD risk factor outcomes were used: hypertension, diabetes, total and high density lipoprotein cholesterol, glycated haemoglobin, fibrinogen, C-reactive protein and Framingham risk score. Four NICE categories of obesity were created based on body mass index (BMI) and waist circumference (WC): no risk (up to normal BMI and low/high WC); increased risk (normal BMI & very high WC, or obese & low WC); high risk (overweight & very high WC, or obese & high WC); and very high risk (obese I & very high WC or obese II/III with any levels of WC. Men and women in the very high risk category had the highest odds ratios (OR) of having unfavourable CVD risk factors compared to those in the no risk category. For example, the OR of having hypertension for those in the very high risk category of the NICE obesity groupings was 2.57 (95% confidence interval 2.06 to 3.21) in men, and 2.15 (1.75 to 2.64) in women. Moreover, a dose-response association between the adiposity groups and most of the CVD risk factors was observed except total cholesterol in men and low HDL in women. Similar results were apparent when the Framingham risk score was the outcome of interest. In conclusion, the current NICE definitions of obesity show utility for a range of CVD risk factors and CVD risk in both men and women [224].
Provide information on resource use and evaluate the potential impact of guidance in changing resource use	There is some scope for undertaking studies around resource use and how this varies by certain patient/service user characteristics. However, all resource use studies (in the absence of establishing data linkages) will be based on self reported data and therefore may lack sufficient depth for some research questions.

Example study and abridged abstract:

"***Inequity and inequality in the use of health care in England: an empirical investigation***: Achieving equity in healthcare, in the form of equal use for equal need, is an objective of many healthcare systems. The evaluation of equity requires value judgements as well as analysis of data. Previous studies are limited in the range of health and supply variables considered but show a pro-poor distribution of general practitioner consultations and inpatient services and a pro-rich distribution of outpatient visits. We investigate inequality and inequity in the use of general practitioner consultations, outpatient visits, day cases and inpatient stays in England with a unique linked data set that combines rich information on the health of individuals and their socio-economic circumstances with information on local supply factors. The data are for the period 1998–2000, just prior to the introduction of a set of National Health Service (NHS) reforms with potential equity implications. We find inequalities in utilisation with respect to income, ethnicity, employment status and education. Low-income individuals and ethnic minorities have lower use of secondary care despite having higher use of primary care. Ward level supply factors affect utilisation and are important for investigating health care inequality. Our results show some evidence of inequity prior to the reforms and provide a baseline against which the effects of the new NHS can be assessed [225].

Provide information on epidemiological trends

HSE data were suggested as being useful to NICE in being able to monitor epidemiological trends, particularly around the social determinants of health. There are several examples across the literature of HSE data being used for these purposes; the example below demonstrates the breadth of HSE in collecting detailed measurements from both adults and children including the collection of salivary data.

Example study and abridged abstract:

"***Recent trends in children's exposure to second-hand smoke in England: cotinine evidence from the Health Survey for England:*** Aims: To examine changes in children's exposure to second-hand tobacco smoke in England since 1998. Design: Repeated cross-sectional

	surveys of the general population in England. The Health Survey for England. A total of 37 038 children participating in surveys from 1998 to 2012, 13 327 of whom were aged 4–15 years, had available cotinine and were confirmed non-smokers. A total of 68.6% (95% CI = 64.3–72.6%) of children had undetectable cotinine in 2012, up from 14.3% (95% CI = 12.7–16.0%) in 1998. There was a highly significant linear trend across years (with a small but significant quadratic term) to declining geometric mean cotinine in all children from 0.52 ng/ml (95% CI = 0.48–0.57) in 1998 to 0.11 ng/ml (95% CI = 0.10–0.12) in 2012. Children from routine/manual backgrounds were more exposed, but experienced similar gains across years to those from non-manual backgrounds. Conclusions: In England, children's exposure to second-hand smoke has declined by 79% since 1998, with continuing progress since smoke-free legislation in 2007. An emerging social norm in England has led to the adoption of smoke-free homes not only when parents are non-smokers, but also when they smoke.
Provide information on current practice to inform the development of NICE quality standards	Some HSE information may be suitable in providing information for the development of NICE quality standards. These data may be particularly useful where the standard is based on meeting a certain level of patient satisfaction or experience, and perhaps less so in providing detailed information on the experience.

Adult Social Care Survey

Aims and description: The Adult Social Care Survey (ASCS) is an annual survey that takes place in England of adult (18+) service user of Local Authority (LA) with Adult Social Services Responsibilities (CASSRS) [49]. The aim of the survey is to collect opinions and experiences of service users on a range of different topics including users' satisfaction with services they received, general health and quality of life and well-being. The survey also allows service users to quality assess the social care provision in their Local Authority. The survey is completed either by the service user or someone who has consented on their behalf (carer, family member).

[49] http://www.hscic.gov.uk/socialcarecollections2015

Background, history and study design:

The ASCS survey is a cross-sectional study developed by the Social Services User Survey Group (SSUSG). This group is a collection of representatives from the Department of Health (DH), the Health and Social Care Information Centre (HSCIC), CASSRs, the Care Quality Commission (CQC), the Association of Directors of Adult Social Services (ADASS) and the Personal Social Services Research Unit (PSSRU). The survey was initially piloted by PSSRU in 2010 and the first survey was conducted during 2010 - 2011.

The survey is tailored for two types of users: residential care or nursing care residents and those receiving community-based services. The sample includes people with a variety of disabilities and problems and receiving a range of services including people with sensory and physical disabilities, learning disabilities (or intellectual impairments), and mental health problems. Services include residential services (e.g. personal care only homes and nursing care homes) as well as community-based services, such as home care and day centres, and other forms of low level or one-off support, such as equipment, transport and meals[226]. The questionnaire is available in a range of different accessible formats (large print, translated, as interview script for face-to-face/phone in addition to a postal survey). The format is generic but each individual LA can include questions that reflect their research needs[50] ; key topics of the survey focus include: satisfaction with services, quality of life, feelings of control, levels and satisfaction with personal care, food and diet provided, accommodation, personal safety, social life, occupation and feelings of dignity.

Validity of measures (e.g. construct, content, criterion validity) and case definitions:

The constructs measured in the survey draw on other survey work and development projects by the SSUSG. Questions included in the survey support ASCOF framework (Adult Social Care Outcome Framework) consisting of seven questions around satisfaction with the quality of care, perceptions of support to maintain independence, understanding of the care system and entitlements, feelings of being in control, feelings of safety and security, and satisfaction with social contact [227]. The survey also includes measures of social care related quality of life using the ASCOT measure, which has been rigorously tested in cognitive interviews [228]. Other measures that have been incorporated into the survey include characteristics of age friendly environments identified by the World Health Organisation [58]. The questionnaire and survey methodology are usually piloted with volunteer LAs before being rolled out nationally [229].

Representative of populations and settings

Individual LAs are requested to send questionnaires to a stratified sample of service users. Respondents receive one reminder, although in residential settings the manager also receives a letter/reminder of completion[226]. Initially the survey was only completed by service users who were receiving complete or partial funding from social services in the same year the survey was carried out; although in 2014-15 the survey was also completed by those whose care needs are assessed and supported through LAs [230]. Data are collected over a defined period across all LAs and a stratified sample approach is used across defined sub-populations; four strata were defined in each LA in 2013/14: all service users

[50] Ibid.p13

with a disability, all other service users aged 18-64, all service users aged 65+ in residential care, and all service users aged 65+ who are resident in the community [229]. Due to the sampling strategy imposed, the results are weighted to ensure that the returned questionnaires are representative of the eligible population. With this scope of research, councils are able to over-sample to produce more robust results based on a specific strata (topic of interest).

While a large number of social care recipients were included in the 2013/14 survey (73,925); the response rate stood at 38 per cent, introducing potential large selection effects in the type of responding user [231]. The weighting strategy accounts only for compositional effects by strata within Local Authorities and does not adjust for the potential impact of non-response. Completion of the survey is now mandatory for LAs providing social care[51], although no additional funding is provided. This means that there is little incentive for investing in improving response rates.

There are concerns around the underrepresentation of certain groups such as adults who have cognitive or memory impairments such as dementia and are likely to be excluded from sampling frames, or when included, are unable to respond [226]. The survey design does not allow for understanding of the full adult social care landscape in England as it either underrepresents or excludes many privately funded service users and those who receive direct payments and purchase their care directly.

Representativeness of different disease stages: While it may be possible to ascertain levels of care needs in the population (as opposed to different disease stages per se), it is not possible to monitor change in needs at an individual level (also see note on consent below).

Clear ethical frameworks: Ethical approval of ASCS has been given by the Social Care Research Ethics Committee; those without capacity to consent to the survey are removed from the sampling frame.

Dynamic and adaptable: The cross-sectional nature of the data does mean that the potential for different epidemiological study designs to be implemented is limited. The survey has adapted to include greater breadth in the eligible population the survey has included changes to its content; there is also scope for LAs to adapt the survey to enable it to better suit their needs.

Data Linkages: There are no clear direct data linkages between ASCS and other data sources.

Collection of data reflective of real-world conditions (scope): The ASCS collects a broad scope of data on user experiences and perceptions. However, there is less scope for understanding how these experiences vary by the intrinsic characteristics of the respondent.

Sensitivity: N/A

[51] Each Local Authority must have an eligible population of 150 or more service users to make it a requirement to participate in the survey, if they have less than 150 users they are not required to participate.

Uniformity in data collection procedures: There is a high degree of uniformity in the surveys in the overall approach. However, there is potential for differences to occur through different policies around fielding surveys in different formats and through LAs having the capacity to insert questions of interest into the survey. These additional questions are submitted to the HSCIC for approval.[52]

Steps taken to minimise common forms of bias: Potential bias could occur through the high levels of non-response to the survey. Forms of bias could also occur in terms of different respondents completing the survey compared to the intended respondents, or respondents receiving assistance to complete the survey [231]; 72% reported receiving assistance in completing the survey. However, allowing proxy respondents to complete or aid completion of the survey enables the responses to better reflect the spectrum of social care clients. Around one-in-ten (9%) intended survey respondents were not involved in completing the survey at all, although this is thought to have minimal impact on the results [231]. As the surveys are conducted by LAs (either directly or contracted out to other agencies) there is scope for deviation to occur in the method of administration; such deviations usually result in the data for some LAs being excluded.

Granularity of treatment/disease data: There is some scope for understanding service users' experiences by their social care needs, measured by the Activities of Daily Living.

Other considerations/information:

More information is needed on the characteristics of eligible populations and a greater understanding is required around the factors that influence non-response rate and how surveys were conducted in each Local Authority [231]. While the results of the survey are published by HSCIC, access to the individual level data may be restricted and the data have not been published on archives such as the UK Data Archive.

Summary and Utility for NICE

ASCS is a survey of users' satisfaction with the care that they receive. Such data can be instrumental in forming guidance that is based on user experience and patient reported outcomes. A disadvantage of the data is that, unlike other surveys of patient or service user experiences, the data are not freely available through sources such as the UK data archive. As such there may be limited scope for undertaking secondary analysis of the individual service user data. Nevertheless, the detailed reports and tables produced still allow for gaining a good level of understanding of aspects of service user satisfaction with their care and broader aspects of wellbeing.

Key References:

HSCIC. Personal Social Services Adult Social Care Survey, England 2013-14, Final release. Leeds: Health and Social Care Information Centre, 2014.

[52] http://www.hscic.gov.uk/article/4793/User-survey-guidance---2014-15, p13

Desired use by NICE	Potential suitability[53]
Research the effectiveness of interventions or practice in real-world (UK) settings	This is a cross-sectional survey of user experiences of social care. While it may be possible to undertake repeated cross-sectional studies and examine the impact of changing practice on user experiences, fully assessing the effectiveness of interventions through measuring longitudinal changes at a service-user level will be challenging with these data.
Audit the implementation of guidance.	It may be possible to assess whether guidance is being implemented, particularly around service user satisfaction or service user reported experiences, through analysing change at a LA level. While the example below does not directly do this, the excerpt highlights that such a study could be possible through examining change over time in reports of being treated with dignity. **Example and excerpts from descriptive report:** A common dictionary definition is *"a state or quality or manner worthy of esteem or respect; and (by extension) self-respect."* while the online practice guide for dignity in care developed in partnership by the Department of Health with the Social Care Institute for Excellence (SCIE) and the Care Services Improvement Partnership (CSIP)[17] offers the following definition *"dignity consists of many overlapping aspects, involving respect, privacy, autonomy and self-worth. The provisional meaning of dignity used in the practice guide is: 'a state, quality or manner worthy of esteem or respect; and (by extension) self-respect'"*. 60 per cent of respondents had a positive reaction to how the way they were helped and treated made them think and feel about themselves in 2013-14 (one percentage point higher than in 2011-12 and 2012-13 and three percentage points higher than in 2010-11). 31 per cent of respondents said the way they were helped and treated did not affect the way they thought and felt about themselves – unchanged from the previous year and down one percentage point from 2010-11 and 2011-12. Eight per cent said it sometimes

[53] Potential suitability here is something of a subjective construct in the absence of a clear research question to be addressed. However,

	undermined the way they thought and felt about themselves and one per cent said it completely undermined them – both of which have remained the same since 2011-12 [231].
Provide information on resource use and evaluate the potential impact of guidance in changing resource use	The data can potentially provide a snapshot of resource use based on service users' reports; repeat cross-sectional studies may provide information on the impact of changing guidance on resource use
Provide information on epidemiological trends	The study provides a snapshot of social care needs but among a population who are receiving LA assistance for these health needs. Given that eligibility for LA provided social care is changing and varies between LAs, there may be some caveats accompanying the results of studies that aim to provide information on how trends vary over time. Nevertheless, there is scope for the information to be used to understand inequalities in health status, wellbeing, or as in the excerpt below, quality of life.
	Example study and extracts from abstract:
	What can local authorities do to improve the social care-related quality of life of older adults living at home? Evidence from the Adult Social Care Survey. This study aims to examine the associations between social care-related quality of life (SCRQoL) in older adults and three potential policy targets for local authorities: (i) accessibility of information and advice, (ii) design of the home and (iii) accessibility of the local area. We used cross-sectional data from the English national Adult Social Care Survey (ASCS) 2010/2011 on service users aged 65 years and older and living at home (N=29,935). To examine the association between SCRQoL, as measured by the ASCOT, and three single-item questions about accessibility of information, design of the home and accessibility of the local area, we estimate linear and quantile regression models. After adjusting for physical and mental health factors and other confounders our findings indicate that SCRQoL is significantly lower for older adults who find it more difficult to find information and advice, for those who report that their home design is inappropriate for their needs and for those who find it more difficult to get

	around their local area. In addition, these three variables are as strongly associated with SCRQoL as physical and mental health factors. We conclude that in seeking to find ways to maintain and improve the quality of life of social care users living at home, local authorities could look more broadly across their responsibilities. Further research is required to explore the cost-effectiveness of these options compared to standard social care services [58].
Provide information on current practice to inform the development of NICE quality standards	There may be scope for the data to be used to form quality standards around social care experiences and trajectories – for example around information advice and guidance received by older people in accessing care. These data may be particularly valuable in providing information to support quality standards aimed at improving people's feelings of dignity when receiving social care

Care.data

Aims and description: Care.data is an initiative which aims to implement a system for GP records data to be shared nationally with the Health and Social Care Information Centre (HSCIC) securely, and for these data then to be linked with other sources in order to create data that can help track patient journeys across the entire continuum of care. In 2013, the HSCIC has described the aims of care.data as being six-fold including: (i) supporting patients' choice; (ii) advance customer service; (iii) promote greater transparency; (iv) improve outcomes; (v) increase accountability; (vi) drive economic growth [232]. If successfully implemented, care.data would make a substantial contribution to the data needs of NICE and other organisations. The research possibilities of such rich data would likely meet many of NICE's needs with respect to real-world data, and would allow for establishing the long-term effectiveness of interventions through the capacity to track patient journeys through primary and into secondary care as standard, something that rarely occurs in real-world data projects and sources. Uniquely, it could also potentially, allow for insight into patterns of social care and their relationship clinical and public health data.

The scheme is being developed first using four Clinical Commissioning Group areas (Blackburn with Darwen, Somerset, West Hampshire and Leeds North, West and South and East) as a test-bed in a pilot (pathfinder) stage. However, at the time of writing, the project progress has faltered; it has been met with widespread concern; and very recently the government's own body on the implementation of major projects voiced concerns about the feasibility of the project [27]. As care.data has not yet been fully implemented at the time of writing, this data source profile does not attempt to recreate a fuller profile provided for other sources of data; instead we focus further on the aims and some of the concerns surrounding the project.

Background, history and study design: The Health and Social Care Act 2012 changed the statutory designation of the Health and Social Care Information Centre (HSCIC) and its roles, responsibilities and powers. One of the changes included that HSCIC can under certain circumstances, including on mandatory request from statutory bodies (NICE being one of the bodies) or when directed by NHS England or the Secretary of State for Health, require individual patient data from GP practices without obtaining explicit additional consent from patients. One of the first examples of this occurring was the request to share data for the care.data project.

Implementation of the care.data project is expected to bring several benefits. Health benefits include improved monitoring of performance and outcomes through linkage between primary and secondary care; earlier diagnoses of illness; predictive modelling of risk and improved decision-making; improved data for evaluation of interventions; and exploration of patient pathways [233]. However, several times an economic argument was also espoused around the commercial value of the data which proved to be unpopular and overlooked the complexities of the debate [234].

The first phases of the scheme were intended to be launched in March 2014. In January 2014, the Guardian newspaper claimed that health data collected under the auspices of care.data would be made available to commercial firms [235]; shortly afterwards HSCIC issued a statement denying that 'data will be made available for the purposes of selling or administering any kind of insurance and that the NHS and the HSCIC never profit from providing data to outside organisations' [54]. However this was later partly refuted as it was revealed that individual patient data (HES data) had been shared with insurance bodies in the past [236]. In the run up to the planned launch and amidst the public controversy, up to 700,000 formally objected to having their data shared with a third party and essentially opted-out. Such a high level of 'non-response' in future impedes the ability of the data to represent a 'census' of NHS records and may mean that weighting strategies may need to be implemented to adjust for differing levels of non-response across groups. It also recently emerged that there are issues in processing this volume of objections and there are technical issues in ensuring that the care of those who raised an objection was not compromised through losing access to services such as screening services[55].

Representativeness (populations and settings; different disease stages):

One of the priority goals of the care.data initiative is to link GP records with other sources of data. Hospital Episodes Statistics (HES) data would become Care Episodes Statistics (CES) data as they would allow for tracing of patient journeys across care settings [233]. Therefore in many ways care.data has the potential to become representative of several different disease stages and to be able to monitor change in disease stage or care need stage in order to facilitate understanding of the effectiveness of care interventions of different forms. However, the status of the 700,000 who have already opted out of care.data is unclear, and therefore an understanding of how the care.data 'population' differs from the wider population is scarce in the literature.

[54] http://www.hscic.gov.uk/article/3869/Response-to-Guardian-article-about-NHS-data-200114
[55] http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/health-committee/handling-of-nhs-patient-data/written/18661.html

The technical specification for the extract of patient data from GP surgeries specified that the following patient-level fields would be included in the data: NHS number (the key linking field); date of birth, gender, postcode, ethnicity, registration details, details of events (e.g. appointments); referral dates and details; information on prescribed medication; information on diagnosis of a range of different conditions (non-communicable diseases); background information on health status (e.g. cholesterol level, smoking status, lung functioning, Q-Risk values); vaccination status; care plans and reviews for different conditions; exception reporting (e.g. where patients had declined advice or treatment). Some data were to be specifically excluded from the extract including HIV status, STI history, abortion history, IVF treatment details, marital status, complaints, convictions, and history of abuse by others [233]. It had been anticipated that care.data would join with other NHS projects including the 100k Genome Project, allowing phenotypic data to be linked with genomic data and enabling advances in precision medicine [28].

Ethical concerns and concerns about anonymity:

In 2014 the project was put on hold, although in 2015 it was announced that the pathfinder scheme would restart, beginning with Blackburn and Darwen. A major factor in the difficulties experienced was a perceived lack of communication with the public about plans to use their health records for medical research. Although NHS England had disseminated an information leaflet through 99% of households, two-thirds of householders did not recall receiving any information about the scheme [237]. This meant that some of the ethical safeguards around informed consent had essentially failed, and some are sceptical about whether this issue has been fully resolved [234]. As discussed earlier, there have also been issues in terms of processing objections and ensuring that these objections are recognised without impacting upon access to services such as screening services.

In their study of care.data, Carter and colleagues identify three main reasons why the lack of social legitimacy of the initiative overruled any legal authority granted in the Health and Social Care Act [238]. Firstly, they identify that there existed defects in the warrants of trust provided for care.data – to this end the levels of awareness were so low that that they rendered the legitimate means of opting out of care.data meaningless. Secondly, care.data ruptured the traditional role and expectations of general practitioners. Thirdly, there emerged widespread uncertainty about the status of care.data as a public good. This was particularly centred on the potential commercialisation of the public's private medical records.

Some have expressed concerns that some of the processes considered for anonymising the data collected (specifically pseudoanonymisation at the source as opposed to at HSCIC) will lead to errors in the data linkage which cannot be rectified centrally. Such inaccuracies are not only likely to increase the risk of clinical error, but will also be detrimental for research and decision-making [239]. However, another concern is around so called 'jigsaw attacks' where data users are able to use the (linked) data released to deliberately identify individuals[56]; while such use of the data is forbidden, there remain concerns that the broad scope of potential users – including academic, public sector,

[56] http://www.england.nhs.uk/wp-content/uploads/2014/01/pia-care-data.pdf

voluntary sector, commercial users in the UK and elsewhere – could equate to weak enforcement of rules of use.

Current status: The Major Projects Authority expressed low confidence in the capacity to deliver care.data and gave the project a red rating in June 2015. A red rating indicates that: "Successful delivery of the project appears to be unachievable. There are major issues on project definition, schedule, budget, quality and/or benefits delivery, which at this stage do not appear to be manageable or resolvable. The project may need re-scoping and/or its overall viability reassessed"[57]. The initiative remains at the 'pathfinder' (pilot) stage and the project is still gathering data on the best ways to support GP practices in terms of the technicalities of sharing data but more importantly in ways that GP surgeries can support patients to make informed choices around their data and its use. It should be noted that this pilot stage does not include piloting the extraction systems themselves or any of the proposed data linkages that would subsequently need to occur to realise the ambitions of the project. Furthermore, the project was not given an end data (one of eight out of 188 major government projects) [27]. A number of actions have been identified that can improve confidence in the implementation of the project; some actions are also reported to have been completed. However, the Major Projects Authority identified that a different rating would be applied when the assurance and approval processes were confirmed, pathfinder stage had been completed and approval for national rollout had been obtained, and the rollout for a nationally linked dataset had been approved; unusually the report also suggested that parts of the business case were still being developed [27]. NHS England have responded that the rating was not current (based on data from September 2014) and that further actions have been completed to the extent that the rating should be downgraded to a less severe category [240].

The Department of Health have also created a National Information Board and have appointed Dame Fiona Caldicott as a national data guardian in efforts to address concerns around confidentiality of data. However, there remain ongoing concerns among academic experts as to the response of bodies responsible for care.data to critiques around the sharing of data with commercial companies [234]. Due to purdah (over spring/summer 2015), there have been few communications around care.data, although work has continued and, subject to approval, extraction of GP records may take place in September 2015 in pathfinder areas only [241]. Should the project be implemented successfully, there are likely to be several opportunities made available to NICE with regards to real-world data usage. However, the degree to which care.data has served to erode public trust in medical research, and particularly medical research involving real-world data, should not be underestimated and may be a factor to consider when exploring suitable sources of real-world data.

It is unfortunate that the controversy surrounding care.data has overshadowed the many potential benefits of seeing the initiative to fruition, which in its full capacity could revolutionise our understanding of patient trajectories, needs and outcomes across the spectrum of care settings that form the mosaic of the health and social care system in the UK. The impact of the negative publicity has served to erode public trust in the use of electronic health records for medical research, while other forms of big data, from store

[57] https://engage.cabinetoffice.gov.uk/major-projects-authority/chapter-2-the-major-projects-authority-remit/

card spending patterns to apps and social network usage, continue to flourish. It has also sparked debate about the very purpose of medical research and widened the perceived gulf between commercial and non-profit research. Nevertheless, it does serve as a reminder that while the technological advances are increasingly permissive towards the incorporation of real-world in health decision-making, there remains a need to understand, maintain and prioritise the ethical frameworks surrounding this real world data collection.

Salford Integrated Record

Aims and description: The Salford Integrated Record (SIR) is an electronic health record that combines patients' primary and secondary health records into a single system. At its core is an aim to ensure that patient records are available to all stakeholders involved in a patient's care, including the patient themselves. The establishment of the record should improve the clinical outcomes of patients through ensuring that clinicians are able to access all relevant patient information in a timely fashion regardless of whether they are treating the patient in primary or secondary care settings [58]. An anonymised research data repository has also been established [242].

Background, history and study design:

The Salford Integrated Record was established in 2001 [243], although information uploaded in the system may pre-date this. Data are updated every 24 hours on patients' GP consultations, hospital episodes, referrals, clinical attendances, as well as on background and selected lifestyle characteristics collected in clinical practice. The SIR also links individual patient prescription information with these primary and secondary data [243]. The SIR was originally focussed on sharing information between providers on four key conditions: diabetes, coronary heart disease, chronic kidney disease and stroke [244].

Validity of measures (e.g. construct, content, criterion validity) and case definitions:

In common with other clinical databases, read codes are used as the basis of data entry. In some cases, read codes can lead to inaccurate estimates of the prevalence of conditions with sets of symptoms that overlap with other conditions. One study used SIR data as the basis of a validation study of the read codes of Irritable Bowel Disease (IBS), finding that levels of IBS were much lower than expected and hypothesised that this was due to a preference for using symptom codes as opposed to disease read codes [245]. The limitation around read code usage is likely to be shared across primary care databases, although the advantage of the SIR is the routine data linkage which could facilitate validation.

Validation studies of coverage have also suggested that SIR is highly concordant with patient information stored on other systems (for COPD and asthma patients), suggesting that SIR offers good levels of valid case definition [97].

Representative of populations and settings: The SIR contains data for GP practices in Salford, Greater Manchester. Fifty-three practices are included and the data are linked with the Salford Royal Hospital. The records of approximately 220,000 residents are

[58] http://www.salfordccg.nhs.uk/download.cfm?doc=docm93jijm4n524.pdf&ver=680

included [246]; all but one GP practice upload their data to the SIR. The data are estimated to account for 97% of Salford's population [96].

Representativeness of different disease stages: The integration of primary and secondary care data allows for research tracking patient outcomes across care providers (through examining Integrated Care Pathways (ICP)). One initiative using the data in this way is the Collaborative Online Care Pathway Investigation Tool, which is being used to examine missed opportunities in patient care – that is where primary prevention opportunities were missed which could lead to adverse health outcomes. This initiative is focussed on modelling the circumstances and frequency of variance between idealised ICP and the actual care provided [16].

Clear ethical frameworks: Each patient has received information on the existence of the record and was given the option of requesting to opt-out of the record (through contacting their GP). Each time the record is accessed by healthcare staff, the permission of the patient needs to be asked before use. The information within the record can be used for medical research purposes, where patient identifiers are removed and medical research ethics committee permission has been granted. The NHS in Salford also advises that research companies may on occasion request patients' personal details, but these are only shared with the express permission of the patient.

Dynamic and adaptable: Several projects are underway that use data from SIR[59]. The potential for longitudinal analysis of patient trajectories allows for a large number of epidemiological study types to be implemented. The data have also been considered for calculations and selection of appropriate outcomes in RCTs and pragmatic trials [246].

Additional patient reported outcomes can also be linked to the data. In the CLASSIC (Comprehensive Longitudinal Assessment of Salford Integrated Care) study for example, older people's (65+) experiences and satisfaction with accessing integrated care will be assessed through a series of questionnaires that are linked to SIR data [60].

Data Linkages:

At its core, the Salford Integrated Record is a data linkage project that links primary and secondary care records to improve patients' clinical outcomes. Linkages are made between local GP practices and the Salford Royal Hospital. Further linkages have been made including with Hospital Episodes Statistics, ONS mortality statistics and the NHS Exeter system (which allows for sharing of data including some screening information, GP payments data and NHS organ donation data) [246]. Pharmacy data have also been linked into the SIR [243].

Collection of data reflective of real-world conditions (scope): The breadth of data collected in the SIR has meant that the data have already been used in effectiveness studies using a pragmatic RCT design [243]. The data have also been used for research questions that can only be addressed using real-world data, such as medication prescribing risks occurring in general practice [247].

[59] http://www.herc.ac.uk/research-development/
[60] http://www.crn.nihr.ac.uk/blog/news/over-4000-people-in-salford-recruited-to-classic-study/

Other considerations/access information: The research potential of the data is being explored by NorthWest e-Health group (NWeH; see http://www.nweh.org.uk/). The availability of the data for researchers and the procedures and costs around accessing the data are unclear from online documentation.

Salford Integrated Record in numbers: 53 GP practices and Salford Royal Hospital; close to 200,000 patient records

Summary and utility for NICE: SIR was suggested as a source of data that may have the potential to overcome the limitations of other data source and examine patients' integrated care pathways. The potential of the data for research purposes are likely to be in the process of being realised and there are comparatively few publications using these data in the literature; the data may have been used primarily to facilitate clinical decision-making and performance management initially. Perhaps one of the most appealing characteristics of the data, given the current climate around the use and ethics of electronic health records in medical research, is the high degree of patient involvement and the ability of patients to access their own records.

Key References:

New JP, Bakerly ND, Leather D, Woodcock A. Obtaining real-world evidence: the Salford Lung Study. *Thorax* 2014:thoraxjnl-2014-205259.

Akbarov A, Kontopantelis E, Sperrin M, Stocks SJ, Williams R, Rodgers S, et al. Primary Care Medication Safety Surveillance with Integrated Primary and Secondary Care Electronic Health Records: A Cross-Sectional Study. *Drug safety* 2015:1-12.

Prescribing Observatory for Mental Health

Aims and description: The Prescribing Observatory for Mental Health (POMH-UK) is a quality improvement programme that is run by the Royal College of Psychiatrists. The aim of the POMH-UK is to work with specialist mental health trusts to improve their prescribing practices. The POMH-UK works with member organisations to identify suitable topics for audit-based quality improvement programmes, the costs of which are borne through subscription costs of member organisations[61].

Background, history and study design:

The POMH-UK was established in 2005 and has been housed at the Centre for Quality Improvement in the Royal College of Psychiatrists since inception [248]. Each quality improvement audit tends focuses on a new topic around prescribing practice in mental health care, although topics tend to be repeated across years. Several audits take place within a year, with different rates of participation among contributing NHS trusts, reflective of the degree of specialism of the given audit. Between June 2006 and October 2013, twelve different topics had been included and thirty-five different audit reports

61 http://www.rcpsych.ac.uk/workinpsychiatry/qualityimprovement/nationalclinicalaudits/prescribingpomh/prescribingobservatorypomh.aspx

produced. The programme tends to start with a baseline audit of practice against evidence-based standards, followed by interventions that aim to change practice and improve adherence to guidance, followed by a re-audit 12-18 months later [249].

Topics covered over this period included: (i) prescribing high dose and combined antipsychotics on adult acute and psychiatric intensive care wards; (ii) screening for metabolic side effects of antipsychotic drugs; (iii) prescribing high dose and combined antipsychotics on forensic wards; (iv) prescribing high dose and combined antipsychotics in cases of acute/PICU, rehabilitation/complex needs, and for forensic psychiatric services; (v) prescribing anti-dementia drugs; (vi) assessment of side effects of depot antipsychotic medication; (vii) monitoring of patients prescribed lithium; (viii) medicines reconciliation; (ix) antipsychotic prescribing in people with learning disabilities; (x) prescribing antipsychotics for children and adolescents; (xi) prescribing antipsychotics in people with dementia; (xii) prescribing for people with personality disorder; and (xiii) prescribing for ADHD in children, adolescents and adults[62]. Topics are chosen on the basis that they meet eight criteria including: that they are relevant to the implementation of particular NICE guidelines; are seen as a clinical priority; where there is likely to be variation at a trust level; and where it is practical and feasible to collect the data[63]. Between 2006 and 2013, the number of cases included in an audit has varied between 1,035 (screening for metabolic side effects of drugs) to 12,790 (prescribing anti-psychotics for people with dementia). The maximum number of participating trusts recorded over this period was 57, with as few as 12 participating in more specialist audits.

In addition to clinical information some additional patient characteristics are collected including the age, gender and ethnicity of the patient [249]. Clinical teams in NHS Mental Health Trusts selected their own audit samples using a variety of methods and can enter as many case histories as they wish [249]. Within trusts, the performance of different clinical teams can be compared.

In recent years, the POMH-UK audits have moved progressively from solely being reliant on case and record submission to audit methods where clinical teams are asked directly about the actions they have taken [250].

Validity of measures (e.g. construct, content, criterion validity) and case definitions:

The criteria used in the audit are expected to be evidence-based and clinical credible (which are not always synonymous [251]).

Representative of populations and settings

Since inception, the number of participating trusts has risen and in addition to NHS trusts, charitable and private organisations also participate in POMH-UK [248]. The results of POMH-UK are generally regarded as generalisable to practice across NHS Mental Health Trusts. While NHS trusts are able to submit as many cases as they want to into the audit, some

62 http://www.rcpsych.ac.uk/workinpsychiatry/qualityimprovement/nationalclinicalaudits/prescribingpomh/prescribingobservatorypomh.aspx

63 http://www.rcpsych.ac.uk/workinpsychiatry/qualityimprovement/nationalclinicalaudits/prescribingpomh/prescribingobservatorypomh/findoutmoreandjoin/pomhqiptopics.aspx

like South London and Maudsley NHS Trust have submitted all cases that fit within the population of interest in previous audits [252].

Representativeness of different disease stages: The data are focussed on different topics at each audit. There may be a greater depth of data collected on disease stage (dependent on the audit) compared to either non-specific databases or databases set up for administrative as opposed to quality assurance purposes [249].

Clear ethical frameworks: Participating trusts are encouraged to contact service users about their involvement in the POMH-UK. Specific patient identifiers such as names and addressed are not shared with POMH-UK.

Data Linkages: There are studies that have used linked NHS trust level data on geographic and healthcare attributes in order to attempt to measure bias in submitted records [251].

Collection of data reflective of real-world conditions (scope): The POMH-UK collects records and case notes from patients. In addition, where necessary, further information reflective of real-world conditions and practice are collected through surveys [253].

Sensitivity: Sensitivity has been used in earlier profiles to denote the ability to distinguish complications from comorbidities. These audit data have been used to understand the emergence of, or more precisely steps taken to prevent and screen for, iatrogenic disease – for example the impact of lithium intake on renal function. However, sensitivity could also indicate the degree to which the data can be used to successfully monitor change, and those working closely with the POMH-UK suggest it can take up to three years to observe a change at a national level in practice, and that any changes are likely to be modest in nature [251].

Uniformity in data collection procedures: While there is a high degree of uniformity in the tools used to conduct audits, there are differences between the way participating trusts select samples.

Steps taken to minimise common forms of bias: Allowing NHS trusts to select their own cases for inclusion into the audit could introduce selection bias in that those cases known to be adhering to the clinical standard of interest are more likely to be included. This form of bias has been examined in studies using the data through analysing linked data based on trust socio-demographic characteristics and its CQC quality ratings to assess whether the submitted samples were over-representative of any group or section of performance [251]. There was little evidence of systematic bias in this assessment. Furthermore, given that participation in the audit is voluntary, creating deliberately biased samples for inclusion in the audit defeats the purpose of participation.

Some of the findings of the audit may reflect poor documentation standards or poor quality of case notes as opposed to actual failure to adhere to guidelines, although the extent of this potential bias is relatively unknown [17].

In the analysis of findings, the central POMH-UK analysts are blinded as to the identity of teams within trusts, and analysis files include codes for trusts, as opposed to names or other identifiers [253].

Granularity of treatment/disease data: ICD-10 disease codes are used to categorise mental health conditions, although the specialist nature of the observatory allows for greater granularity of psychiatric morbidity [249].

Summary and utility for NICE

One of the key criteria for choosing a topic focus of the POMH-UK is that the topics are relevant for monitoring the implementation of NICE guidelines. This has direct relevance to one of the intended uses of real-world data by NICE. An example of study directly assessing the implementation of NICE guidance can be found in a study of renal and thyroid functioning among patients who are prescribed lithium [17].

However, these audit data are very much focussed on comparing practice against pre-defined standards; therefore as emphasised by Paton and Barnes [251], the utility of the data for other more research-focused or evaluative activities, for example in assessing the effectiveness of interventions or monitoring epidemiological trends, may consequently be limited. The data are not widely used in the literature and it is not clear the extent to which these data are available for re-analysis of individual patient data, reflecting their primary function as a quality improvement tool. Nevertheless, there are several important questions that could be addressed for NICE as there may be potential to understand not only whether practice/outputs changed over time but also the processes that led to these changes, as highlighted in Mace and Taylor's account of change in one NHS Trust [252].

Key References:

Barnes TRE, Paton C. Improving prescribing practice in psychiatry: The experience of the Prescribing Observatory for Mental Health (POMH-UK). *International Review of Psychiatry* 2011;23(4):328-335.

Section 5: Summary of in-depth profiles of data and organisational case study

Section 4 includes detailed profiles of 11 data sources that were primarily named in interviews as being useful for NICE. Here we include a summary of these sources and we also provide information collected from an organisation with a similar profile to NICE on the way in which real-world data contributes to their practice.

Selection of findings

Primary care databases

- Three primary care databases are included in the in-depth data profiles: THIN, QResearch and CPRD. All three differ slightly in coverage and include between 587-950 GP practices, and achieve a level of coverage of between 6-9 per cent (THIN being the smallest and QResearch the largest). There has been little work to try to understand the similarities or differences of these data in the literature. Each data source has a similar approach in that they are reliant on using the information inputted into GP systems as part of clinical care for research purposes; no additional data is routinely collected/inputted (although other that may be linked) apart from the data that is inputted as a result of GP consultations, vaccinations, blood tests and similar data (for example data from registration questionnaires). There may be provision for collecting additional data, however. THIN and CPRD are most closely structured in terms of the underlying databases, and have similar access procedures. They also have overlapping GP surgeries – the same GP surgery can appear in both datasets. QResearch is the largest of the databases but uses different software to input the data. All three databases are reliant on the voluntary cooperation of GPs (as well as of patients who can drop out of the data) although validation studies suggest that all three are broadly representative with some minor variances.
- All three have identical systems of recording information (read codes) and there is a substantial degree of uniformity in this respect. However, none of the data sources is infallible to problems of misclassification. There are also shared weaknesses with respect to data quality and breadth of sociodemographic and socioeconomic data. Some minor differences are present. QResearch data will not include free-text data whereas CPRD and THIN do; THIN and CPRD also explicitly state that additional data – e.g. patient reported outcomes – can be collected although this is not explicit/a feature in QResearch documentation.
- All three datasets have been used to conduct sophisticated analyses and could be used to implement studies that meet all five of NICE's needs of real world data. However, QResearch, being the newest of the three sources, appears less frequently in the literature. It may also be less advanced in terms of data linkages, although this is changing very rapidly. CPRD is notable in having rebranded to reflect the intention to link primary care data with hospital episodes and other data. Studies that use linked data for any of the three main primary care datasets

are in the minority and are comparatively rare. CPRD data appears more frequently in the literature, either as CPRD or its forerunner the GPRD. THIN and CPRD have markedly different access policies to QResearch; data in QResearch is only available to research consortiums where the PI is based in a UK university; THIN and CPRD access policies are less bounded. All three have a cost affixed to accessing the data that varies by the size, breadth and complexity of the data needed and the type of provider.

Table 3: UK Primary Care databases in numbers

	Number of GP practices	Number of patient records	Number of active patients
CPRD	685	13.7 million	4.4 million
QResearch	950	13 million	~5.1 million[64]
THIN	587	12.4 million	3.7 million

Data on resource use and quality - utilising service user reports

- Two of the datasets profiled in section 5 – the Adult Social Care Survey and the Community Mental Health Survey - highlight a relatively untapped resource around monitoring the way in which guidelines, and particularly those aimed at improving patient experiences, can be used to monitor changes at a population level before and after implementation. One of the issues with these specific data is that changes in the population included or in the questions asked means that comparability across recent sweeps has been impeded. Nevertheless, these data do offer a potential resource to monitor such changes in future.
- A number of survey data sources including those above as well as the Health Survey for England and the English Longitudinal Study for Ageing allow for monitoring changes in resource use through patient/service user reports. While the data may only be suitable for providing headline trends and may lack some of the detail required for assessing resource use, they may nevertheless present useful sources of data, particularly given the breadth of other data on the characteristics of patients/service users. Furthermore the latter data source allows for assessing patterns of longitudinal change among individuals as well as population level change.

Data linkage studies remain underrepresented and under-utilised at a national level

- Some real-world data sources are purposefully either set-up or re-developed to enhance their data linkages and to examine the presence/absence of integrated patient care. This is a key area for real world data if it is to meet the evidence needs of decision-makers and clinicians and to keep pace with policy

[64] England only: http://www.qresearch.org/PowerPointpresentations/Validity%20and%20completeness%20of%20the%20NHS%20Number%20in%20%20primary%20and%20secondary%20care%20data%20in%20England%201991-2013.pdf

developments. However, those that are designed to enable the monitoring of care across providers, or at least have the capability to do so at a national level, have been utilised relatively rarely for this purpose. Furthermore, none of the data sources we have profiled have been used in studies that have examined transitions between clinical and social care, despite the increasing recognition of the interdependency between these sectors. One of our interviewees did however consider the most likely candidate where this type of data would occur was in either local data sources, such as the Salford Integrated Record or in registry or audit data. In the latter case the more focussed nature of the research questions being considered may necessitate these data to be linked and for integrated care pathways across the continuum of care to become a data attribute. On a national level, care.data may provide a catalyst for this type of data to become available for health researchers, although there are several milestones to reach before this becomes a viable possibility.

Using survey data to understand the feasibility and impact of guidance (and potential support needs)

- Among the data sources we examined in the in-depth data profiles, it was clear that no one data source represented a panacea for NICE's real world data needs. Where some data sources would give detailed information on the nature of the clinical or social care episode, there was less information available on the characteristics of the patient or service users. Similarly, where there was detailed information on the socio-demographic and lifestyle characteristics of patients or service users, the information on clinical or social care interactions would be less granular. This does highlight the merits and importance of data linkage projects succeeding although as discussed above, these types of studies are generally underrepresented. The findings from the in-depth profiles are suggestive of a need to triangulate evidence across different data, particularly in order to understand the feasibility and impact of guidance. For example, while detailed granular information from large clinical databases may be useful for setting standards and developing guidance around GP consultations (for example), information from surveys can help to reveal the feasibility of this guidance on a patient level and illuminate where inequalities may lie in simple measures of GP access. We highlighted earlier that survey data was generally underutilised within NICE; these in-depth profiles illuminate the large number of different study designs and research questions that can be addressed using these data.

Using clinical and administrative data for medical research

- These in-depth profiles of data sources highlight that the research potential of some is yet to be fully realised. For example, many of the data sources focussed on service user experiences and those focussed on auditing practice rarely appear in the peer-reviewed literature. However, this does not diminish either their robustness of the data or their future potential. NICE's own intended uses of real-world data cover a spectrum of needs from addressing in-depth research questions around the effectiveness and variance in the effectiveness of interventions to more audit purposes. A broad brushed approach to considering different forms of real-world data is likely to yield substantial benefits.

Addressing concerns around data quality and bias

- The findings highlight that researchers have used a variety of methods to investigate and address potential sources of bias in the data. These include various sensitivity analyses and analytical techniques to investigate and limit the impact of sources of bias including confounding by indication and selection bias. Other forms of bias around measurement have been addressed through implementing rigorous quality assurance processes and fielding instruments that have been rigorously tested across other data sources. Similarly, the sensitivity of some of the sources have been exploited to examine the occurrence of iatrogenic diseases over a long exposure time, and even to detect changes in practice following the issuing of NICE guidelines. A good deal of familiarity with the data source is likely to be required to understand the extent of potential bias and implement techniques that may help to address these. Nevertheless, all the sources profiled in-depth all show a substantial degree of potential to address the real world data needs of NICE where the appropriate analytical technique is employed. Furthermore, should a similar profile have been developed for all the named data sources in section 2, we expect that the same conclusions would have been drawn.

Addressing gaps in coverage

- The in-depth profiles revealed that some data sources – notably the CPRD and THIN – hold the potential for additional data to be collected, particularly around Patient Reported Outcomes. Furthermore, some of the data sources were specifically focussed on understanding patients' experiences and satisfaction with services, while others collected in-depth measures of patients' mental wellbeing which could be associated with their health state (although do not fall within the remit of 'patient reported outcomes' strictly in terms of outcomes following clinical or care interventions). These in-depth profiles offer ideas of how patient and service user input could be incorporated to a greater extent into NICE's work. Some of the gaps identified earlier (section 2) do remain, even after creating the in-depth data profiles. These include obtaining data on care home or extra care housing residents who are privately funded as well as data on service user experience of public health services, for example user experiences and characteristics in sexual health programmes or smoking cessation programmes. There may also be gaps in the covering different disease stages; for example end of life/palliative care was not a focus in this review but some sources, for example the 2013 sweep of the Health Survey for England, have included modules covering this issue.

Focus on the potential utility of different datasets for NICE

All the profiled data sources are likely to have some utility to NICE dependent on the research question and making a specific recommendation around use is challenging as this is very much dependent on the context and the focus of the research question. The following section summarises the utility of the different sources for NICE.

English Longitudinal Study of Ageing (ELSA)

- ELSA has been used to establish the effectiveness of interventions at a population level using observational methods, for example in a cost-benefit analysis of cataract surgery among ELSA respondents [4]. ELSA may be less suitable for establishing the effectiveness of more specialist interventions/practice, or establishing how interventions/practice vary among minority groups.
- ELSA can be used to determine the implementation of guidance through examining broad population-level temporal changes in the receipt of common interventions or practice. For example ELSA data were used to examine shortfalls in care for chronic conditions using set quality indicators [5]. Without further linkages, ELSA data may be less suitable for explaining the underlying mechanisms around the implementation of guidance, beyond patient/service-user characteristics.
- ELSA data can be used to provide information on some aspects of resource use for example how many people receive common interventions and can be used to establish how access may vary by individual patient characteristics.
- ELSA data can be used to establish self-reported levels and determinants of many age related conditions and non-communicable diseases and more broadly information on lifestyle behaviours and attitudes among older people.
- ELSA data may be less suitable for establishing the incidence/prevalence/outcomes of very uncommon diseases/conditions/interventions.

Community Mental Health Survey

- The CMHS data have been used to monitor the implementation of guidance, for example in monitoring the implementation of guidance aiming to strengthen support for service users during times of turnover in staffing [6]. The data have also been used to draw together guidance around expected standards of care [7]. There may also be potential to use the data to monitor different aspects of resource usage.
- The focus of the survey is on service user experiences and there is less information on outcomes following receipt of different forms of care, limiting the utility of the data with respect to establishing the effectiveness of interventions. The data are less suitable as a tool for monitoring epidemiological patterns in mental health.

Clinical Practice Research Datalink (CPRD)

- CPRD data have utility for NICE through the flexibility in being able to collect additional fields. CPRD data are also available to medical researchers based outside UK universities potentially expanding the pool of potential partners with which NICE could work in using the dataset. The long established nature of CPRD (based on

GPRD) means that several retrospective studies could also be potentially conducted using these data.
- There are numerous examples where CPRD (and GPRD) data have been used in studies that cover all of NICE's intended uses of real-world data. For example, CPRD have been used to evaluate changes in cancer diagnostic intervals following the introduction of NICE guidance [8]. Given the potential to draw large samples, studies could be implemented that examine the epidemiology/outcomes/implementation of rare or less common conditions and procedures. Unlike survey-based sources, for example ELSA and HSE, and in the absence of further data collection, there is potential to examine only a limited range of patient-level intrinsic factors, although these may be sufficient for many studies.
- Data linkages will expand the utility of CPRD data for NICE; current linkages include those with MINAP data, National Cancer Intelligence Network data and HES data. Area level data are also available including Index of Multiple Deprivation data and Townsend deprivation scores [9]. Further data linkages are planned.

QResearch

- QResearch is of interest to NICE for many of the real world data uses identified by NICE, but access appears to be restricted to research consortiums led by academic institutions. Nevertheless, given the substantial potential of these data, NICE could consider ways of developing research projects based on QResearch data led by universities.
- There is potential for QResearch data to be used in studies that cover all of NICE's intended uses of real-world data. The use of QResearch data in developing risk prediction scores may also be of interest to NICE, potentially around forecasting and modelling future disease burden.
- Given the potential to draw large samples, studies can be implemented that examine the epidemiology/outcomes/implementation of rare or less common conditions and procedures. One example is a study of peanut allergy, where a prevalence rate of 0.51 per 1000 patients in the UK was estimated [10].
- The study depositors state that QResearch data are suitable for case control studies designed to examine risk factors for onset of disease, cross sectional surveys, cohort studies and sample size calculations (for non-observational studies) [11].
- As is the case for all three large primary care databases, there is potential to examine only a limited range of patient-level intrinsic factors, although these may be sufficient for many studies.

The Health Improvement Network (THIN)

- There are numerous examples where THIN data have been used in studies that cover all of NICE's intended uses of real-world data. For example, THIN data have been used to examine equity in access to cancer screening among people with Intellectual Disabilities compared to those without across different types of cancer [12].
- Data linkages expand the utility of THIN, and THIN data have been linked with Hospital Episodes Statistics (HES) data, providing potential for studying continuity in care between primary and secondary care. A number of patient postcode-based socioeconomic, ethnicity and environmental indicators are available to researchers including Townsend deprivation quintile scores.

- Overall there is a wide scope for analysing data reflecting outcomes and experiences of morbidity and mortality at primary care level, as well as trends in the care and treatment provided. These data can also be linked to HES data allowing for potential tracking of patient journeys between primary and secondary care. As is the case for all three large primary care databases, there is potential to examine only a limited range of patient-level intrinsic factors, although these may be sufficient for many studies.
- THIN data have utility for NICE through the flexibility in being able to collect additional fields and the potential to conduct research based on free-text fields. THIN data are also available to medical researchers based outside UK universities potentially expanding the pool of potential partners with which NICE could work with in utilising real world data.

National Minimum Data Set for Social Care (NMDS-SC)

- NMDS-SC is a specialist dataset suitable for monitoring trends in the social care workforce. This data can potentially help NICE to understand workforce capabilities and undertake preliminary work to understand the feasibility of implementing new standards and guidance in social care settings.
- The data may be suitable to examine changes following the implementation of NICE guidance at a workforce level in terms of indicators such as pay, training or necessary skills. They may also be useful in helping to set benchmarks and develop quality standards around workforce capacity and skills. The data have also been incorporated into calculations of resource use in the literature [13]. While the data do not provide insight into epidemiological trends per se, they do provide insight into the workforce preparedness for responding to epidemiological challenges, such as dementia [14].
- As social care outcomes are not collected in NMDS-SC, it is unlikely that these data are suitable for researching the effectiveness of interventions and practice.

Health Survey for England (HSE)

- HSE was suggested in the context of monitoring epidemiological trends although the potential usage extends beyond this purpose alone and potentially HSE data can be used to gain an understanding of trends over time in terms of resource utilisation, trends in social care needs and usage, trends in lifestyles and social determinants of health, and some trends in prescribing, service usage and attitudes to health. With regards to researching the effectiveness of interventions, in the absence of data linkages, there may be more limited potential to measure the effectiveness of interventions or changes in practice. Examples where data have been linked to explore later outcomes include an examination of fruit and vegetable intake and mortality [15].
- The survey data may be of great utility for NICE in gathering contextual information critical in the assessing feasibility of different forms of guidance aimed at public health and social care challenges. The data also have the added advantage of being relatively easy to obtain for further secondary data analysis and are free to use.
- There is scope for auditing the implementation of guidance through examining change in practice at a population level; one of the strengths of HSE data in doing so is the ability to examine social or medical inequalities in the implementation of

guidance. Some HSE information may be suitable in providing information for the development of NICE quality standards and these data may be particularly useful where the standard is based on meeting a certain level of patient satisfaction or experience.

Adult Social Care Survey

- ASCS is a survey of users' satisfaction with the care that they receive. Such data can be used in forming guidance that is based on user experience and patient reported outcomes. There may be limited scope for undertaking secondary analysis of the individual service user data without further permissions being sought. Nevertheless, the detailed reports and tables produced may allow for gaining a good level of understanding of aspects of service user satisfaction with their care and broader aspects of wellbeing.
- With regards to measuring the effectiveness of practice, while it may be possible to undertake repeated cross-sectional studies and examine the impact of changing practice on user experiences, fully assessing the effectiveness of interventions through measuring longitudinal changes at a service-user level will be challenging with these data. However, it may be possible to assess whether guidance is being implemented, particularly around service user satisfaction or service user reported experiences, through analysing change (for example at a Local Authority level).
- With regards to using the data as an epidemiological tool, the study provides a snapshot of general health trends and social care needs but among a population who are receiving LA assistance for these health needs (the sample design represents a caveat around the applicability of the data). There may be scope for the data to be used to form quality standards around social care experiences and trajectories – for example around information advice and guidance received by older people in accessing care.

Salford Integrated Record

- SIR was suggested as a source of data that may have the potential to overcome the limitations of other data source and examine patients' integrated care pathways. The potential of the data for research purposes are likely to be in the process of being realised and there are comparatively few publications using these data in the literature; the data may have been used initially to mainly facilitate clinical decision-making and performance management. Perhaps one of the most appealing characteristics of the data, given the current climate around the use and ethics of electronic health records in medical research, is the high degree of patient involvement and the ability of patients to access their own records.
- The data hold substantial potential for improving patient care. The integration of primary and secondary care data allows for research tracking patient outcomes across care providers (through examining Integrated Care Pathways (ICP)). One initiative using the data in this way is the Collaborative Online Care Pathway Investigation Tool that is being used to examine missed opportunities in patient care – that is where primary prevention opportunities were missed which could lead to adverse health outcomes. This initiative is focussed on modelling the circumstances and frequency of variance between idealised ICP and the actual care provided [16].

Prescribing Observatory for Mental Health

- One of the key criteria for choosing a topic focus of the POMH-UK is that the topics are relevant for monitoring the implementation of NICE guidelines. This has direct relevance to one of the intended uses of real-world data by NICE. An example of study directly assessing the implementation of NICE guidance can be found in a study of renal and thyroid functioning among patients who are prescribed lithium [17].
- The utility of the data for other more research-focused or evaluative activities, for example in assessing the effectiveness of interventions or monitoring epidemiological trends, may be more limited. The data are not widely used in the literature and it is unclear the extent to which these data are made available for re-analysis, reflecting their primary function as a quality improvement tool. Nevertheless, there are several important questions that could be addressed for NICE as there may be potential to understand whether practice/outputs have changed over time. In addition, this source represents one of the few specialist sources of real-world data on mental health encountered.

Care.data

- If successfully implemented, care.data would make a substantial contribution to the real-world data needs of NICE and other organisations. The data could allow for establishing the long-term effectiveness of interventions through the capacity to track patient journeys through primary and into secondary care as standard, something that rarely occurs as standard in real-world data projects and sources. Uniquely, it could also potentially, allow for insight into patterns of social care and their relationship clinical and public health data.
- At the time of writing it is too early to tell the extent to which care.data has been able to overcome the challenges encountered, particularly around consent and conditions around data usage. The results of the pathfinder exercise will offer further insight into the viability of the whole project; the majority of testing in pathfinder areas is due to begin later this year.

Using real-world data in Healthcare Technology Assessment: experiences from the Swedish Council on Healthcare Technology Assessment (SBU)

Background

Interviewees were asked for their input to help identify organisations with a similar remit to NICE who were using real-world data in an extensive and effective way. One of the criteria for selecting an organisation for a case study was that the organisation's experience of real-world data should be primarily as a user of real-world data and not as a producer. Some suggestions were received, although most promising in terms of comparability were suggestions of examining the way in which the extensive sources of real-world data in Sweden were used by organisations with a similar remit to NICE.

Section 5: Summary of in-depth profiles of data and organisational case study - Using real-world data in Healthcare Technology Assessment: experiences from the Swedish Council on Healthcare Technology Assessment (SBU)

Context of Real-world data in Sweden

Sweden has an extensive framework of central government funded registers which encompass aspects of clinical healthcare, public health and social care. These are often characterised by having extensive coverage, scope, are available for medical research, and are operated on a non-commercial basis. One example is the National Patient Register, a complete census of all inpatient events since 1987 and a census of inpatient and outpatient events across private and public hospitals (including psychiatric hospitals) since 2006[65]. Another is the Swedish Pharmaceuticals Registry, which provides information on prescribed medications and is compulsory for all pharmacies operating in the country[66]. In addition to national, centrally funded registries, there is also an extensive network of more localised or more focused registers that cover specific diseases/conditions. A recent review by Emilsson and colleagues assessed 103 of these active registers (known as Quality Registries) [254]. These registers are intended to improve the delivery of patient outcomes by monitoring quality standards and adherence to guidelines, investigating disparities in healthcare by geographic and social characteristics, and comparing the effectiveness of different interventions. This network of specialist registers complements the coverage of national registers through providing the depth of information not found in more generic national clinical databases. It is this breadth of potential sources of real-world data that makes Sweden an interesting context for examining real-world data use.

Context of SBU [and TLV; TBC]

SBU is one of the oldest Health Technology Assessment (HTA) organisations in the world, established in 1987. SBU take a broad approach to assessing interventions and consider the clinical, social, economic and ethical dimensions of interventions. The organisation primarily relies on systematic reviews as its prime research methodology and has made methodological contributions in the systematic review field[67]. Unlike NICE, it has a narrower remit in terms of its decision-making functions. Its prime function is to present comprehensive, rigorous and impartial summaries of the evidence around interventions and to present these findings to decision-making authorities such as the National Board of Health and Welfare, the Medical Products Agency and the Dental and Pharmaceutical Benefits Agency. Therefore, an assessment of the use of real-world data in HTA needs to incorporate the experience of both SBU and another decision-making authority. This case study focuses on the experiences and viewpoints of SBU and TLV (Dental and Pharmaceutical Benefits Agency [TBC]) in real-world data against a context of a rich landscape of real-world data sources.

SBU experience of using real-world data

SBU currently use real-world data primarily to understand the use of the evidence it produces. Of NICE's intended data usages; this corresponds closely with the aim of auditing the implementation of guidance.

Most of the organisation's current data usage is based on evidence from RCTs (and in some cases evidence from observational studies), used as part of systematic reviews. This is

[65] http://www.socialstyrelsen.se/register/halsodataregister/patientregistret/inenglish
[66] http://www.jpi-dataproject.eu/Home/Database/388?topicId=1
[67] http://www.sbu.se/en/About-SBU/

rarely accessed in the form of granular individual patient data. Where real world data is used is around the assessment of whether SBU's reports and conclusions are used in healthcare decision-making, as opposed to using real-world data to create these reports. However, the organisation is planning to incorporate real-world data within more aspects of its work and there is substantial appetite and interest within the organisation to do so. SBU has been developing relationships in order to realise these ambitions.

The specific aspects of interest for SBU in using real-world data are around establishing the generalisability of evidence from RCT studies and examining the effectiveness of interventions using the extensive registry data collected in Sweden. SBU is currently discussing with partner organisations how to develop methods to understand the utility of these data.

Currently, registry data is rarely used in SBU's work, although summaries of evidence of register-based studies are likely to feature in the systematic reviews it collects. The most common form of real-world data in use are based on surveys of practitioners and decision-makers where SBU aims to assess the degree to which its findings are influencing clinical practice. The surveys are administered in-house and usually receive a remarkably high response rate (>90%). The questions are usually limited to assessing the impact of findings, and the results are used to understand how to better improve the production and dissemination of SBU's findings, and the results are published on the organisation's website and in discussion and debate papers. Unlike NICE's intended use of real-world data, SBU has not yet used real-world data at any point in making definitive recommendations, primarily because the remit of the organisation differs from that of NICE.

Nevertheless, there is recognition that there is potential for much greater usage and particularly in incorporating the wealth of new data sources established locally in the past decade. While SBU has been able to present evidence on the safety of interventions, the use of real-world data in future will be able to help it present evidence around the effectiveness of interventions, and there is particular interest in sources of real-world data that allow for the administration of pragmatic trials. While the organisation is at an early stage in expanding its profile of real-world data use, one of the important considerations that will feature in its quality assurance assessments of real-world data sources is the population coverage. Many of Sweden's national registers have over 90 per cent population coverage, allowing for understanding of local patterns of healthcare delivery. As much of Sweden's healthcare system is devolved at the municipal level, a data source that allows for understanding of municipal level trends will be an important consideration in appraising sources of real-world data. As the trend in the English NHS is also following suit, this may be an increasingly important shared criteria in assessing sources of real-world data.

Key Findings and Recommendations

Key Findings

- The real-world data landscape remains complex and heterogeneous and composed of sources with different purposes, structures and collection methods. This heterogeneity may increase with opportunities stemming from the incorporation of new technologies in data collection (current quality assured sources are limited in number)
- Some real-world data sources are purposefully either set-up or re-developed to enhance their data linkages and to examine the presence/absence/effectiveness of integrated patient care; however, such sources are in the minority. Furthermore, the small number that are designed to enable the monitoring of care across providers, or at least have the capability to do so at a national level, have been utilised infrequently for this purpose in the literature.
- Data that offer the capacity to monitor transitions between health and social care do not currently exist at a national level, despite the increasing recognition of the interdependency between these sectors.
- Among the data sources we included, it was clear that no one data source represented a panacea for NICE's real world data needs. This does highlight the merits and importance of data linkage projects and is suggestive of a need to triangulate evidence across different data, particularly in order to understand the feasibility and impact of guidance.

Key Overall Recommendation

- There exists no overall catalogue or repository of real-world data sources for health, public health and social care, and previous initiatives aimed at creating such a resource have not been maintained. As much as there is a need for enhanced usage of the data, there is also a need for taking stock, integration, standardisation, and quality assurance of different sources. This research highlights a pressing need for a systematic approach to creating an inventory of sources with detailed meta-data and the funding to maintain this resource. This would represent an essential first step to support future initiatives aimed at enhancing the use of real-world data.

Key Recommendations for NICE

Increased utilisation of existing sources beyond clinical databases:

- Making recommendations is difficult around the use of specific data sources. However, NICE's current use of real-world data differs substantially from the landscape with respect to its low utilisation of clinical audit, disease registry and survey data. Several of the datasets profiled in-depth highlight the potential of different sources of survey, clinical database and audit data.
- We also recommend that NICE further review its use of disease registry and audit data and engage in dialogue with collectors and depositors of these data to explore the utility of these types of data. Sources, such as those available from

the National Cancer Data Repository, are currently underutilised or entirely overlooked during the production of guidance or technology assessments.

Investment in capacity and partnership building

- Use of real-world data requires substantial investment of resource that allows for the organisation to develop an in-depth understanding and experience of using different real-world sources. The extent of this undertaking should not be underestimated; any commitments and real-world data usage strategies should be matched by resources that allow for developing expertise in-house and in developing partnerships with study depositors and academic experts.
- Many of the data sources profiled either have active user groups or hold regular consultative exercises. NICE should further investigate these opportunities and capitalise on these.

Strategy and influence

- NICE has the potential to influence the availability of real-world data sources and good practice around the collection and utilisation of real-world data. This influence could be used to develop good practice around aspects such as obtaining informed consent from patients or obtaining investment around the creation of data linkages. NICE should develop and publish an outward-facing policy around its use of real-world data which includes transparent means of influencing the state of the landscape, in order to ensure that sources continue to meet its organisational needs. Exerting such influence could not only lead to benefits to NICE, but will have broader positive impacts across other stakeholders more widely, and could lead to improved patient and service user outcomes. This influence could also extend to developing quality standards around the way in which data are collected that can be shared across the sector.
- Care.data represents an initiative that could potentially meet many of NICE's real-world data needs. NICE should engage in discussions with HSCIC to better understand and prepare for potentially using these data, while continuing to monitor whether and how the initiative overcomes challenges identified in earlier stages.

Understanding implementation

- Finally, while NICE is potentially able to monitor the implementation of guidelines using several sources, it may still lack information on the underlying mechanisms as to how or why guidelines succeed or fail in implementation. Starting its own programme of real-world data collection in the form of surveys of practitioners may be a way of understanding the mechanisms of un/successful implementation. Such an approach has been adopted elsewhere, for example by the Swedish Council on Healthcare Technology Assessment (SBU).

Green shoots

There are three key factors as to why the state of the real world data landscape should be regarded with some optimism for NICE and more generally.

1. Firstly, while data linkage and the capacity to research patient journeys is not at the point where many would desire, there are several examples where these efforts have been met with success and some of these have been met with a high degree of public acceptance. On a national level, the care.data initiative has restarted after a pause, and if these efforts succeed, they could meet many of NICE's real-world data requirements.

2. Secondly, while we have been critical in the study about the representation of sources of patient reported outcomes, there are examples featured in the main report where patients have become more involved and have become gatekeepers to their own data (e.g. Salford Integrated Record), providing a possible model for the future. In addition, the ubiquity of smartphone technology and apps mean that ways of patients providing and managing their own information are increasing at pace.

3. Thirdly, methodological advances in the design and analysis of studies continue to ensure that real-world data becomes of greater utility for organisations, such as NICE, who wish to understand the implications of their decisions in real-world settings. These advances include the development of pragmatic trials using electronic health data which offer a balance between the methodological rigour of RCTs and the generalisability of observational studies. Several UK based organisations and teams – some of which are represented among the expert stakeholders involved in the present study – are involved in driving these advances and it is likely that future studies will feature the results of these undertakings extensively in their findings.

Glossary

Real-world data

The definition of real-world data can be contentious and different stakeholders have different views as to what constitutes 'real-world' data. Real world data is defined in this report through two key tenets:

a. The collection of real world data reflects the usual care or treatment provided to populations of patients, service users or the public. This therefore excludes conventional Randomised Controlled Trials (RCT(s)) but could include other forms of RCT design, namely pragmatic RCTs.

b. Real world data provides enough depth to assess trends around everyday practice, service usage, or assess outcomes.

Clinical databases

Clinical databases usually collect data from a particular form of service or a set of services. They have undefined/multiple entry points and less well-defined criteria around case definitions than is the case for registry data (and consequently have no defined denominator[68]). Due to the breadth of data collected, they can be used to address a variety of research questions although they do vary in their quality and value in addressing some questions (e.g. some forms of epidemiological surveillance). Multicentre databases are particularly useful tools for monitoring/assessing population health. Clinical databases may be particularly useful for monitoring conditions that may be difficult to diagnose or where there are no definitive clinical tests but where diagnoses are made through observing symptomology (using a structured checklist) and/or response to medication; they may be less useful for monitoring very rare conditions.

Disease/case registries

Disease registers are effectively a census of all cases of a particular disease or health condition [3 105], and can additionally include much more detailed accounts of treatment regimes. Disease registries can be population based, hospital based or clinical based; or can reflect some other form of community boundary. Detailed case histories can lend disease registry data particularly suitable for some forms of research (for example pharmaco-epidemiological studies and resource use studies). Disease registries are viewed as important tools in improving (i) patient care; (ii) public health; (iii) technology assessment and (iv) information provision [105]. Disease registers are most useful in situations where disease or risk factor status does not tend to change over time; the diagnosis of disease needs to be consistent, and based on a robust diagnostic test. Furthermore, a register is of use when there is a requirement for ongoing health care, for example, retinal screening among patients with diabetes [2].

[68] Although clinical databases are often intended to be representative of a specific population with a denominator in mind

Workforce registers/case registers

Workforce or case registers are another form of real-world data. They can include data on qualifications, training and continuous professional development, specialism, and socio-demographic information, and provide indicative evidence around resource use and skill surpluses and deficits. They can be collected by public health authorities as well as representative bodies who may require staff to be registered in order to practice.

Healthcare technology/surgical registers

Surgical/technology registers are similar in design to disease registers in that they effectively comprise a census of all specific surgical procedures or technological implementation and the outcomes of these procedures. Pharmacoepdiemiological registers/databases comprise another form of register, although are rarely confined to a single drug and are more akin to clinical databases in having a broader case definition.

Surveys

Data provided by surveys can help identify specific problems in the delivery of healthcare services or the health status of individuals. Surveys can provide a much deeper understanding of the lifestyles and antecedent factors surrounding contact with health or social care providers; longitudinal (e.g. panel or cohort studies) provide an opportunity to develop models of causal inference. Surveys can also be useful for determining patients' views about the care that they receive as well as other factors surrounding patients and service users including their mental health and quality of life outcomes. There can exist methodological issues in the design of surveys with how patients are sampled, non-response and missing data which can compromise the generalisability of results [2]. In addition, sample size can be a substantial limitation in general purpose surveys where the number of observed contacts with service providers and/or observed episodes of ill health or care needs can be relatively low.

Clinical audits

A clinical audit is a process that has been defined by NICE as "a quality improvement process that seeks to improve patient care and outcomes through systematic review of care against explicit criteria and the implementation of change" [255]. In many ways the distinction between clinical audits and disease registries may be fuzzy in terms of design, although the remit of disease registries can be broader in that they can be focussed upon predetermined scientific (research) purposes as well as clinical and policy purposes; clinical audits are arguably more focused on performance against standards. Generally, it is accepted that there are few distinguishing features between clinical audits and registries set-up for research purposes [256], although in practice some differences may arise in terms of the scale of collection, structure and establishment and funding of data, and some of these differences have, on occasion, limited the use (or dissemination) of clinical audit data for research purposes [257].

Population registries and censuses

Another common form of real-world data are population registries and censuses. The census has included questions on self-rated health and caring, and a proportion of census respondents are also monitored more closely in some studies (e.g the LS study; see Appendix 2). Included in this category are those registers that capture vital events such as

fertility and mortality; the universal coverage of vital statistics registers mean that they comprise censuses of births and death for large geographic areas.

References

1. van Staa T-P, Goldacre B, Gulliford M, et al. Pragmatic randomised trials using routine electronic health records: putting them to the test. Bmj 2012;**344**:e55.
2. Gnani S, Majeed A. *A user's guide to data collected in primary care in England*. Cambridge: Eastern Region Public Health Observatory, 2006.
3. Newton J, Garner S. Disease registers in England. Oxford: Institute of Health Sciences, University of Oxford, 2002.
4. Weale M. A cost-benefit analysis of cataract surgery based on the English Longitudinal Survey of Ageing. Journal of health economics 2011;**30**(4):730-39.
5. Steel N, Bachmann M, Maisey S, et al. Self reported receipt of care consistent with 32 quality indicators: national population survey of adults aged 50 or more in England. Bmj 2008;**337**.
6. CQC. National Summary of the Results for the 2014 Community Mental Health Survey. London: Care Quality Commission, 2014.
7. Kendall T, Crawford MJ, Taylor C, et al. Improving the experience of care for adults using NHS mental health services: summary of NICE guidance. Bmj 2012;**344**.
8. Neal RD, Din NU, Hamilton W, et al. Comparison of cancer diagnostic intervals before and after implementation of NICE guidelines: analysis of data from the UK General Practice Research Database. British journal of cancer 2014;**110**(3):584-92.
9. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). International journal of epidemiology 2015:dyv098.
10. Kotz D, Simpson CR, Sheikh A. Incidence, prevalence, and trends of general practitioner-recorded diagnosis of peanut allergy in England, 2001 to 2005. Journal of Allergy and Clinical Immunology 2011;**127**(3):623-30. e1.
11. QResearch. What is QResearch? Secondary What is QResearch? 2012. http://www.qresearch.org/SitePages/What%20Is%20QResearch.aspx.
12. Osborn DPJ, Horsfall L, Hassiotis A, et al. Access to cancer screening in people with learning disabilities in the UK: cohort study in the health improvement network, a primary care research database. 2012.
13. Curtis L. Unit Costs of Health and Social Care. Canterbury, Kent: PSSRU, University of Kent, 2009.
14. Hussein S, Manthorpe J. The dementia social care workforce in England: secondary analysis of a national workforce dataset. Aging & mental health 2012;**16**(1):110-18.
15. Oyebode O, Gordon-Dseagu V, Walker A, et al. Fruit and vegetable consumption and all-cause, cancer and CVD mortality: analysis of Health Survey for England data. Journal of epidemiology and community health 2014:jech-2013-203500.
16. Ainsworth J, Buchan I. COCPIT: a tool for integrated care pathway variance analysis. Studies in health technology and informatics 2011;**180**:995-99.
17. Collins N, Barnes TRE, Shingleton-Smith A, et al. Standards of lithium monitoring in mental health trusts in the UK. BMC psychiatry 2010;**10**(1):80.
18. Black N, Barker M, Payne M. Cross sectional survey of multicentre clinical databases in the United Kingdom. Bmj 2004;**328**(7454):1478.
19. Raftery J, Roderick P, Stevens A. Potential use of routine databases in health technology assessment. Health Technology Assessment 2005;**9**(20):1-106.
20. Morabia A. Observations Made Upon the Bills of Mortality. Bmj 2013;**346**.
21. Collins R. What makes UK Biobank special? The Lancet 2012;**379**(9822):1173-74.
22. Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. New England Journal of Medicine 2012;**366**(6):489-91.
23. Okun S, McGraw D, Stang P, et al. Making the case for continuous learning from routinely collected data. Washington, DC: Institute of Medicine, 2013.

24. Yiu C. The big data opportunity: making government faster, smarter and more personal. London: Policy Exchange, 2012.
25. BMA. Care.data confidentiality concerns cannot be ignored, say doctors. Secondary Care.data confidentiality concerns cannot be ignored, say doctors 2014. http://bma.org.uk/news-views-analysis/news/2014/march/caredata-confidentiality-concerns-cannot-be-ignored-say-doctors.
26. Goldacre B. Care.data is in chaos. It breaks my heart *The Guardian* 2014 28-02-2014.
27. Cabinet Office. Major Projects Authority Annual Report 2014-15. London: Major Projects Authority, 2015.
28. Nuffield Council on Bioethics. The collection, linking and use of data in biomedical research and health care: Ethical issues. London: Nuffield Council on Bioethics, 2015.
29. Devlin N, Appleby J. Getting the most out of PROMs: Putting health outcomes at the heart of NHS decision-making. London: King's Fund, 2010.
30. Miani C, Robin E, Horvath V, et al. Health and Healthcare: Assessing the Real World Data Policy Landscape in Europe. Santa Monica: RAND Corporation, 2014.
31. King D, Wittenberg R. Data on Adult Social Care: Report of School for Social Care Research Scoping Study. London: School for Social Care Research, National Institute for Health Research, 2015.
32. Humphries R. Our response to the proposed new partnership for health and social care in Greater Manchester. *King's Fund* 2015.
33. McCall B. UK medical research gets political. The Lancet 2015;**385**(9976):1381-83.
34. Hussein S. The use of'large scale datasets' in UK social care research. London: NIHR School for Social Care Research, London School of Economics and Political Science, 2011.
35. Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to bisphosphonates and risk of gastrointestinal cancers: series of nested case-control studies with QResearch and CPRD data. BMJ: British Medical Journal 2013;**346**.
36. Keltie K, Cole H, Arber M, et al. Recommendations to NICE on the use of routine data sources in evidence development for NICE IP Guidance. NICE Medical Technologies Evaluation Programme (MTEP). Newcastle-Upon-Tyne: NICE, 2015.
37. ABPI. Demonstrating Value with Real World Data. London: The Association of the British Pharmaceutical Industry, 2011.
38. Morrato EH, Elias M, Gericke CA. Using population-based routine data for evidence-based health policy decisions: lessons from three examples of setting and evaluating national health policy in Australia, the UK and the USA. Journal of Public Health 2007;**29**(4):463-71.
39. Harris KM, Kneale D, Lasserson TJ, et al. School-based self management interventions for asthma in children and adolescents: a mixed methods systematic review. The Cochrane Library 2015(4).
40. Ferreira-González I, Marsal JR, Mitjavila F, et al. Patient Registries of Acute Coronary Syndrome Assessing or Biasing the Clinical Real World Data? Circulation: Cardiovascular Quality and Outcomes 2009:CIRCOUTCOMES. 108.844399.
41. Hill EM, Turner EL, Martin RM, et al. "Let's get the best quality research we can"• : public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study. BMC Medical Research Methodology 2013;**13**(1):72.
42. John A, Dennis M, Kosnes L, et al. Suicide Information Database-Cymru: a protocol for a population-based, routinely collected data linkage study to explore risks and patterns of healthcare contact prior to suicide to identify opportunities for intervention. BMJ open 2014;**4**(11):e006780.
43. Kane R, Wellings K, Free C, et al. Uses of routine data sets in the evaluation of health promotion interventions: opportunities and limitations. Health Education 2000;**100**(1):33-41.

44. Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to statins and risk of common cancers: a series of nested case-control studies. BMC cancer 2011;**11**(1):409.
45. Smeeth L, Douglas I, Hall AJ, et al. Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials. British journal of clinical pharmacology 2009;**67**(1):99-109.
46. Alderwick H, Ham C, Buck D. Population Health Systems: Going beyond integrated care. London: King's Fund, 2015.
47. Williams T, Van Staa T, Puri S, et al. Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. Therapeutic Advances in Drug Safety 2012;**3**(2):89-99.
48. Keltie K, Cole H, Arber M, et al. Identifying complications of interventional procedures from UK routine healthcare databases: a systematic search for methods using clinical codes. BMC Medical Research Methodology 2014;**14**(1):126.
49. Hennekens CH, Buring JE. *Epidemiology in medicine*. Philadelphia: Lippincott Williams & Wilkins, 1987.
50. Gliklich RE, Dreyer NA, Leavy MB. *Registries for evaluating patient outcomes: a user's guide*. Rockville, MD: Agency for Healthcare Research and Quality (US), Government Printing Office, 2014.
51. Freemantle N, Marston L, Walters K, et al. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. Bmj 2013;**347**.
52. NICE. Headaches: Diagnosis and management of headaches in young people and adults. London: National Institute for Health and Care Excellence, 2012.
53. NICE. Review of TA223; Cilostazol, naftidrofuryl oxalate, pentoxifylline and inositol nicotinate for the treatment of intermittent claudication in people with peripheral arterial disease. London: National Institute for Health and Care Excellence, 2014.
54. Sheldon TA, Cullum N, Dawson D, et al. What's the evidence that NICE guidance has been implemented? Results from a national evaluation using time series analysis, audit of patients' notes, and interviews. Bmj 2004;**329**(7473):999.
55. Platt C, Larcombe J, Dudley J, et al. Implementation of NICE guidance on urinary tract infections in children in primary and secondary care. Acta Paediatrica 2015.
56. Tugnet N, Pearce F, Tosounidou S, et al. To what extent is NICE guidance on the management of rheumatoid arthritis in adults being implemented in clinical practice? A regional survey. Clinical Medicine 2013;**13**(1):42-46.
57. NICE. Inflammatory bowel disease Quality Standard. London: NICE, 2015.
58. van Leeuwen KM, Malley J, Bosmans JE, et al. What can local authorities do to improve the social care-related quality of life of older adults living at home? Evidence from the Adult Social Care Survey. Health & place 2014;**29**:104-13.
59. Skills for Care. National Minimum Data Set for Social Care. Secondary National Minimum Data Set for Social Care.
60. Hussein S, Manthorpe J, Stevens M. Social care as first work experience in England: a secondary analysis of the profile of a national sample of migrant workers. Health & social care in the community 2011;**19**(1):89-97.
61. Blak BT, Thompson M, Dattani H, et al. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Informatics in primary care 2012;**19**(4):251-55.
62. Springate DA, Kontopantelis E, Ashcroft DM, et al. Clinical Codes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. PloS one 2014;**9**(6):e99825.
63. QResearch. What is QRESEARCH? Secondary What is QRESEARCH? 2012.
64. Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: a new general practice database for research. Informatics in primary care 2004;**12**(1):49-50.
65. Collins GS, Altman DG. Identifying patients with undetected pancreatic cancer in primary care: an independent and external validation of QCancer®(Pancreas). British Journal of General Practice 2013;**63**(614):e636-e42.

66. Hippisley-Cox J. QResearch. Primary Health Care Specialist Group Conference. Stratford-upon-Avon, Warwickshire, 2014.
67. Coloma PM, Schuemie MJ, Trifirò G, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiology and drug safety 2011;**20**(1):1-11.
68. Quint JK, Mallerova H, DiSantostefano RL, et al. Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). BMJ open 2013;**4**(7):e005540.
69. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. British Journal of General Practice 2010;**60**(572):e128-e36.
70. THIN. The Health Improvement Network. Secondary The Health Improvement Network 2015.
71. Taggar JS, Coleman T, Lewis S, et al. The impact of the Quality and Outcomes Framework (QOF) on the recording of smoking targets in primary care medical records: cross-sectional analyses from The Health Improvement Network (THIN) database. BMC public health 2012;**12**(1):329.
72. Marmot MG, Stansfeld S, Patel C, et al. Health inequalities among British civil servants: the Whitehall II study. The Lancet 1991;**337**(8754):1387-93.
73. Singh-Manoux A, Kivimaki M, Glymour MM, et al. Timing of onset of cognitive decline: results from Whitehall II prospective cohort study. Bmj 2012;**344**.
74. Bouillon K, Singh-Manoux A, Jokela M, et al. Decline in low-density lipoprotein cholesterol concentration: lipid-lowering drugs, diet, or physical activity? Evidence from the Whitehall II study. Heart 2011:hrt. 2010.216309.
75. NICOR. MINAP. Secondary MINAP 2014.
76. Herrett E, Smeeth L, Walker L, et al. The myocardial ischaemia national audit project (MINAP). Heart (Bmj) 2010;**96**(16):1264.
77. Birkhead JS, Weston CFM, Chen R. Determinants and outcomes of coronary angiography after non-ST-segment elevation myocardial infarction. A cohort study of the Myocardial Ischaemia National Audit Project (MINAP). Heart 2009;**95**(19):1593-99.
78. Bhaskaran K, Hajat S, Haines A, et al. Short term effects of temperature on risk of myocardial infarction in England and Wales: time series regression analysis of the Myocardial Ischaemia National Audit Project (MINAP) registry. Bmj 2010;**341**.
79. Mindell J, Biddulph JP, Hirani V, et al. Cohort profile: the health survey for England. International journal of epidemiology 2010;**41**(6):1585-93.
80. Geoffroy MC, Hertzman C, Li L, et al. Morning salivary cortisol and cognitive function in mid-life: evidence from a population-based birth cohort. Psychological medicine 2012;**42**(08):1763-73.
81. Rahman S, Ecob R, Costello H, et al. Hearing in 44-45 year olds with m. 1555A> G, a genetic mutation predisposing to aminoglycoside-induced deafness: a population based cohort study. BMJ open 2012;**2**(1):e000411.
82. Knies G, Burton J. Analysis of four studies in a comparative framework reveals: health linkage consent rates on British cohort studies higher than on UK household panel surveys. BMC Medical Research Methodology 2014;**14**(1):125.
83. Stevens KN, Lang IA, Guralnik JM, et al. Epidemiology of balance and dizziness in a national population: findings from the English Longitudinal Study of Ageing. Age and ageing 2008;**37**(3):300-05.
84. Brewer M, Browne J, Emmerson C, et al. Pensioner poverty over the next decade: what role for tax and benefit reform? London: Institute for Fiscal Studies, 2007.
85. Buck N, McFall S. Understanding Society: design overview. Longitudinal and Life Course Studies 2011;**3**(1):5-17.
86. Knies G. UK Household Longitudinal Study: Wave 1-4, 2009-2013 User Manual. Colchester: UK Data Service, 2014.

87. McAloney K, Graham H, Hall J, et al. OP13 Diet and physical activity levels among UK youth. Journal of epidemiology and community health 2012;**66**(Suppl 1):A6-A6.
88. Shiue I, Hristova K. Associated social factors of hypertension in adults and the very old: UK Understanding Society cohort, 2009-2010. International journal of cardiology 2013;**168**(4):4563-65.
89. Halpin L, Savulescu J, Talbot K, et al. Improving access to medicines: empowering patients in the quest to improve treatment for rare lethal diseases. Journal of medical ethics 2013:medethics-2013-101427.
90. Springbett A. Scottish Health Informatics Programme (SHIP). Scottish Medicines Consortium Event. Edinburgh, 2011.
91. Livingstone SJ, Looker HC, Hothersall EJ, et al. Risk of cardiovascular disease and total mortality in adults with type 1 diabetes: Scottish registry linkage study. PLoS medicine 2011;**9**(10):e1001321.
92. Anwar H, Fischbacher CM, Leese GP, et al. Assessment of the under-reporting of diabetes in hospital admission data: a study from the Scottish Diabetes Research Network Epidemiology Group. Diabetic Medicine 2011;**28**(12):1514-19.
93. Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. International journal of epidemiology 2013;**42**(1):97-110.
94. Boyd A, Golding J, Macleod J, et al. Cohort profile: the 'children of the 90s': the index offspring of the Avon Longitudinal Study of Parents and Children. International journal of epidemiology 2012:dys064.
95. ALSPAC. The Avon Longitudinal Study of Parents and Children - ALSPAC. Bristol: University of Bristol, 2014.
96. Quantifying the longitudinal value of healthcare record collections for pharmacoepidemiology. AMIA Annual Symposium Proceedings; 2011. American Medical Informatics Association.
97. Elkhenini HF, Davis KJ, Stein ND, et al. Using an electronic medical record (EMR) to conduct clinical trials: Salford Lung Study feasibility. BMC medical informatics and decision making 2015;**15**(1):8.
98. Kavanagh SJ, Bray B, Paley L, et al. Abstract W P288: Using 'Big Data'• Analytics and Visualization for Quality Improvement in Stroke Care. Stroke 2015;**46**(Suppl 1):AWP288-AWP88.
99. Harrison DA, Welch CA, Eddleston JM. The epidemiology of severe sepsis in England, Wales and Northern Ireland, 1996 to 2004: secondary analysis of a high quality clinical database, the ICNARC Case Mix Programme Database. Crit Care 2006;**10**(2):R42.
100. Nolan JP, Laver SR, Welch CA, et al. Outcome following admission to UK intensive care units after cardiac arrest: a secondary analysis of the ICNARC Case Mix Programme Database*. Anaesthesia 2007;**62**(12):1207-16.
101. Kemp K, O. UK IBD Audit Steering Group. OC-094 UK inflamatory bowel disease audit: nurse correlations between 2006 and 2008. Gut 2010;**59**(Suppl 1):A39.
102. Pokhrel S, Owen L, Lester-George A, et al. Tobacco Control Return on Investment Tool. London: National Institute for Health and Care Excellence. London: National Institute for Health and Care Excellence, 2014.
103. NICE. Assessing service levels for cardiovascular disease prevention. NICE commissioning guides [CMG45]: Services for the prevention of cardiovascular disease. London: National Institute for Health and Care Excellence, 2012.
104. Kristensen FB, Palmhøj Nielsen C, Chase D, et al. What is Health Technology Assessment? In: Velasco Garrido M, Kristensen FB, Palmhøj Nielsen C, et al., eds. Health Technology Assessment and Health Policy-Making in Europe Current status, challenges and potential. Copenhagen, Denmark: WHO Regional Office for Europe, 2008.
105. Rankin J, Best K. Disease registers in England. Paediatrics and Child Health 2014;**24**(8):337-42.

106. Herrett E, Thomas SL, Schoonen WM, et al. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. British journal of clinical pharmacology 2010;**69**(1):4-14.
107. PHE. National Cancer Intelligence Network Cancer statistics: availability and location. London Public Health England, 2014.
108. Challis D, Clarkson P, Warburton R. *Performance indicators in social care for older people*: Ashgate Publishing, Ltd., 2006.
109. Kneale D, Smith L. Extra Care Housing in the UK: Can it be a Home for Life? Journal of Housing for the Elderly 2013;**27**(3):276-98.
110. Miller C, Bunnin A, Rayner V. Older people who self fund their social care: A guide for health and wellbeing boards and commissioners London: OPM, 2013.
111. Steptoe A, Breeze E, Banks J, et al. Cohort profile: the English longitudinal study of ageing. International journal of epidemiology 2013;**42**:1640–48.
112. Wiggins RD, Netuveli G, Hyde M, et al. The evaluation of a self-enumerated scale of quality of life (CASP-19) in the context of research on ageing: A combination of exploratory and confirmatory approaches. Social Indicators Research 2008;**89**(1):61-77.
113. Zivin K, Llewellyn DJ, Lang IA, et al. Depression among older adults in the United States and England. The American Journal of Geriatric Psychiatry 2010;**18**(11):1036-44.
114. Hamer M, Lavoie KL, Bacon SL. Taking up physical activity in later life and healthy ageing: the English longitudinal study of ageing. British journal of sports medicine 2013:bjsports-2013-092993.
115. Zaninotto P. Gender differences in quality of life and depression among older people with coronary heart disease. UCL (University College London), 2012.
116. Pierce MB, Zaninotto P, Steel N, et al. Undiagnosed diabetes-data from the English longitudinal study of ageing. Diabetic Medicine 2009;**26**(7):679-85.
117. Zaninotto P, Jackson S, Jackowska M, et al. 4. Trends in obesity among older people in England. In: Banks J, Nazroo J, Steptoe A, eds. The Dynamics of Ageing Evidence from the English Longitudinal Study of Ageing 2002 – 2012. London: Institute of Fiscal Studies, 2014:94.
118. Bridges S, Hussey D, Blake M, et al. 5. Methodology. In: Banks J, Nazroo J, Steptoe A, eds. The Dynamics of Ageing Evidence from the English Longitudinal Study of Ageing 2002 – 2012. London: Institute of Fiscal Studies, 2014:94.
119. Chan KS, Kasper JD, Brandt J, et al. Measurement equivalence in ADL and IADL difficulty across international surveys of aging: findings from the HRS, SHARE, and ELSA. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 2012;**67**(1):121-32.
120. Kneale D. Connected communities? LGB older people and their risk of exclusion from decent housing and neighbourhoods. Quality in Ageing and Older People forthcoming.
121. Steptoe A, Breeze E, Banks J, et al. Cohort profile: the English longitudinal study of ageing. International journal of epidemiology 2013;**42**(6):1640–48.
122. Weir D, Faul J, Langa K. Proxy interviews and bias in the distribution of cognitive abilities due to non-response in longitudinal studies: a comparison of HRS and ELSA. Longitudinal and Life Course Studies 2011;**2**(2):170.
123. NatCen. English Longitudinal Study of Ageing (ELSA): User Guide to the Datasets. London: National Centre for Social Research, 2012.
124. Gray M. A review of Data linkage procedures at NatCen London: NatCen, 2009.
125. Kobayashi LC, Wardle J, von Wagner C. Limited health literacy is a barrier to colorectal cancer screening in England: evidence from the English Longitudinal Study of Ageing. Preventive Medicine 2014;**61**:100-05.
126. Rippon I, Kneale D, de Oliveira C, et al. Perceived age discrimination in older adults. Age and ageing 2014;**43**(3):379-86.

127. Judge A, Welton NJ, Sandhu J, et al. Equity in access to total joint replacement of the hip and knee in England: cross sectional study. Bmj 2010;**341**.
128. Bridges S, Hussey D, Blake M. The dynamics of ageing: The 2012 English Longitudinal Study of Ageing (Wave 6) Techincal Report. London: NatCen, 2015.
129. Bowling A, Windsor J. The effects of question order and response-choice on self-rated health status in the English Longitudinal Study of Ageing (ELSA). Journal of epidemiology and community health 2008;**62**(1):81-85.
130. Gale CR, Cooper C, Sayer AA. Prevalence of frailty and disability: findings from the English Longitudinal Study of Ageing. Age and ageing 2015;**44**(1):162-65.
131. The impact of primary care supply on quality of care in England. Health & Healthcare in America: From Economics to Policy; 2014. Ashecon.
132. Picker Institute Europe. Guidance Manual for the NHS Community Mental Health Service Users Survey 2013` Oxford: Picker Institute Europe, 2013.
133. Picker Institute Europe. Guidance Manual for the NHS Community Mental Health Service Users Survey 2014 Oxford: Picker Institute Europe, 2014.
134. Picker Institute Europe. Community Mental Health Survey 2014: Sampling Errors. Oxford: Picker Institute Europe, 2014.
135. MHRA. General Practice Research Database. London: Medicines and Healthcare products Regulatory Agency, 2001.
136. CPRD. Observational Data. Secondary Observational Data 2015. http://www.cprd.com/ObservationalData/CodedData.asp#ObservationalText.
137. Mathur R, Grundy E, Smeeth L. Availability and use of UK based ethnicity data for health research: NCRM: National Centre for Research Methods, 2013.
138. Reeves D, Springate DA, Ashcroft DM, et al. Can analyses of electronic patient records be independently and externally validated? The effect of statins on the mortality of patients with ischaemic heart disease: a cohort study with nested case-control analysis. BMJ open 2014;**4**(4):e004952.
139. Rañopa M, Douglas I, Staa T, et al. The identification of incident cancers in UK primary care databases: a systematic review. Pharmacoepidemiology and drug safety 2015;**24**(1):11-18.
140. UK Terminology Centre. SNOMED CT: A user guide for General Practice. Leeds: UK Terminology Centre, 2012.
141. Herrett E, Shah AD, Boggon R, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. Bmj 2013;**346**:f2350.
142. Bhaskaran K, Forbes HJ, Douglas I, et al. Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). BMJ open 2013;**3**(9):e003389.
143. Mathur R, Bhaskaran K, Chaturvedi N, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. Journal of Public Health 2014;**36**(4):684-92.
144. Bhatnagar P, Wickramasinghe K, Williams J, et al. The epidemiology of cardiovascular disease in the UK 2014. Heart 2015:heartjnl-2015-307516.
145. Brown I, Brown L, Korff D. Using NHS patient data for research without consent. Law, Innovation and Technology 2010;**2**(2):219-58.
146. O'Meara H, Carr DF, Evely J, et al. Electronic health records for biological sample collection: feasibility study of statin-induced myopathy using the Clinical Practice Research Datalink. British journal of clinical pharmacology 2014;**77**(5):831-38.
147. McDonald HI, Nitsch D, Millett ERC, et al. New estimates of the burden of acute community-acquired infections among older people with diabetes mellitus: a retrospective cohort study using linked electronic health records. Diabetic Medicine 2014;**31**(5):606-14.
148. Pfeil AM, Imfeld P, Pettengell R, et al. Trends in incidence and medical resource utilisation in patients with chronic lymphocytic leukaemia: insights from the UK

Clinical Practice Research Datalink (CPRD). Annals of hematology 2015;**94**(3):421-29.
149. Hughes G, Martinez C, Myon E, et al. The impact of a diagnosis of fibromyalgia on health care resource use by primary care patients in the UK: an observational study based on clinical practice. Arthritis & Rheumatism 2006;**54**(1):177-83.
150. de Vries F, De Vries C, Cooper C, et al. Reanalysis of two studies with contrasting results on the association between statin use and fracture risk: the General Practice Research Database. International journal of epidemiology 2006;**35**(5):1301-08.
151. Walker AJ, Card T, Bates TE, et al. Tricyclic antidepressants and the incidence of certain cancers: a study using the GPRD. British journal of cancer 2011;**104**(1):193-97.
152. Brookhart MA, Stürmer T, Glynn RJ, et al. Confounding control in healthcare database research: challenges and potential approaches. Medical care 2010;**48**(6 0):S114.
153. Schneider-Lindner V, Delaney JA, Dial S, et al. Antimicrobial drugs and community-acquired methicillin-resistant Staphylococcus aureus, United Kingdom. Emerging infectious diseases 2007;**13**(7):994.
154. Amirthalingam G, Andrews N, Campbell H, et al. Effectiveness of maternal pertussis vaccination in England: an observational study. The Lancet 2014;**384**(9953):1521-28.
155. Holden SE, Jenkins-Jones S, Poole CD, et al. The prevalence and incidence, resource use and financial costs of treating people with attention deficit/hyperactivity disorder (ADHD) in the United Kingdom (1998 to 2010). 2013.
156. Myles PR, McKeever TM, Pogson Z, et al. The incidence of pneumonia using data from a computerized general practice database. Epidemiology and infection 2009;**137**(05):709-16.
157. Edwards CJ, Campbell J, van Staa T, et al. Regional and temporal variation in the treatment of rheumatoid arthritis across the UK: a descriptive register-based cohort study. BMJ open 2012;**2**(6):e001603.
158. Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: a new general practice database for research. Journal of Innovation in Health Informatics 2004;**12**(1):49-50.
159. QResearch. A summary of public health indicators using electronic data from primary care Nottingham: QResearch and the Health and Social Care Information Centre, 2008.
160. Hippisley-Cox J, Fenty J, Heaps M. Trends in Consultation Rates in General Practice 1995 to 2006: Analysis of the QRESEARCH database. London: QResearch and The Information Centre for health and social care, 2007.
161. Vinogradova Y, Coupland C, Hippisley-Cox J. Use of combined oral contraceptives and risk of venous thromboembolism: nested case-control studies using the QResearch and CPRD databases. Bmj 2015;**350**:h2135.
162. Simpson CR, Hippisley-Cox J, Sheikh A. Trends in the epidemiology of smoking recorded in UK general practice. British Journal of General Practice 2008;**60**(572):e121-e27.
163. Hippisley-Cox J. Validity and completeness of the NHS Number in primary and secondary care: electronic data in England 1991-2013. Nottingham: University of Nottingham, 2013.
164. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. Bmj 2007;**335**(7611):136.
165. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. Heart 2008;**94**(1):34-39.
166. Hippisley-Cox J, Coupland C. Unintended effects of statins in men and women in England and Wales: population based cohort study using the QResearch database. Bmj 2010;**340**.

167. Hippisley-Cox J, Vinogradova Y. Trends in consultation rates in general practice 1995/1996 to 2008/2009: analysis of the QResearch database. London: Health and Social Care Information Centre, 2009.
168. Hippisley-Cox J, Vinogradova Y. Trends in Consultation Rates in General Practice 1995 to 2008: Analysis of the QResearch® database. Leeds: NHS Information Centre, 2009.
169. Simpson CR, Hippisley-Cox J, Sheikh A. Trends in the epidemiology of chronic obstructive pulmonary disease in England: a national study of 51 804 patients. British Journal of General Practice 2010;**60**(576):e277-e84.
170. IMS Health. IMS Health About Us, 2015.
171. Lewis JD, Schinnar R, Bilker WB, et al. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. Pharmacoepidemiology and drug safety 2007;**16**(4):393-401.
172. Blak BT, Thompson M, Dattani H, et al. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Informatics in primary care 2011;**19**(4):251-55.
173. Cai B, Xu W, Bortnichak E, et al. An algorithm to identify medical practices common to both the General Practice Research Database and The Health Improvement Network database. Pharmacoepidemiology and drug safety 2012;**21**(7):770-74.
174. UCL THIN Research. The THIN database. Secondary The THIN database 2015. https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database.
175. Seminara NM, Abuabara K, Shin DB, et al. Validity of The Health Improvement Network (THIN) for the study of psoriasis. British Journal of Dermatology 2011;**164**(3):602-09.
176. Ruigómez A, Martín-Merino E, Rodríguez LAG. Validation of ischemic cerebrovascular diagnoses in the health improvement network (THIN). Pharmacoepidemiology and drug safety 2010;**19**(6):579-85.
177. Denburg MR, Haynes K, Shults J, et al. Validation of The Health Improvement Network (THIN) database for epidemiologic studies of chronic kidney disease. Pharmacoepidemiology and drug safety 2011;**20**(11):1138-49.
178. Re VL, Haynes K, Forde KA, et al. Validity of The Health Improvement Network (THIN) for epidemiologic studies of hepatitis C virus infection. Pharmacoepidemiology and drug safety 2009;**18**(9):807.
179. Langley TE, Szatkowski L, Gibson J, et al. Validation of The Health Improvement Network (THIN) primary care database for monitoring prescriptions for smoking cessation medications. Pharmacoepidemiology and drug safety 2010;**19**(6):586-90.
180. Marston L, Carpenter JR, Walters KR, et al. Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: a cross-sectional study. BMJ open 2014;**4**(4):e004958.
181. Powell HA, Iyen-Omofoman B, Baldwin DR, et al. S93 COPD and risk of lung cancer: The importance of smoking and timing of diagnosis of COPD. Thorax 2012;**67**(Suppl 2):A46-A46.
182. Joseph A. One-time consent patient data for research has already been granted ethical approval; Response to Wendler, D (2006) One-time general consent for research on biological samples. BMJ 2006;**332**:544.
183. Petersen I, Gilbert RE, Evans SJ, et al. Pregnancy as a major determinant for discontinuation of antidepressants: an analysis of data from The Health Improvement Network. The Journal of clinical psychiatry 2011;**72**(7):979-85.
184. Smith CJP, Gribbin J, Challen KB, et al. The impact of the 2004 NICE guideline and 2003 General Medical Services contract on COPD in primary care in the UK. QJM 2008;**101**(2):145-53.
185. Fardet L, Petersen I, Nazareth I. Risk of cardiovascular events in people prescribed glucocorticoids with iatrogenic Cushing's syndrome: cohort study. Bmj 2012;**345**:e4928.

186. González-Pérez A, Gaist D, Wallander M-A, et al. Mortality after hemorrhagic stroke Data from general practice (The Health Improvement Network). Neurology 2013;**81**(6):559-65.
187. Rodríguez LAG, Tolosa LB. Risk of upper gastrointestinal complications among users of traditional NSAIDs and COXIBs in the general population. Gastroenterology 2007;**132**(2):498-506.
188. Toh S, Rodríguez LAG, Hernández-Díaz S. Use of antidepressants and risk of lung cancer. Cancer Causes & Control 2007;**18**(10):1055-64.
189. Guest JF, Panca M, Sladkevicius E, et al. Clinical outcomes and cost-effectiveness of continuous positive airway pressure to manage obstructive sleep apnea in patients with type 2 diabetes in the UK. Diabetes care 2014;**37**(5):1263-71.
190. Wijlaars L, Nazareth I, Petersen I. Trends in depression and antidepressant prescribing in children and adolescents: a cohort study in The Health Improvement Network (THIN). PloS one 2012;**7**(3):e33181.
191. Skills for Care. NMDS-SC key information and statistics reports: May 2015. Leeds: Skills for Care, 2015.
192. Skills for Care. NMDS-SC Raw Data User Guide - Worker File. Leeds: Skills for Care, 2013.
193. Hussein S, Manthorpe J. Structural marginalisation among the long-term care workforce in England: evidence from mixed-effect models of national pay data. Ageing and Society 2014;**34**(01):21-41.
194. Skills for Care. NMDS-SC Return from local authorities and Data Protection. Secondary NMDS-SC Return from local authorities and Data Protection 2013. https://www.nmds-sc-online.org.uk/help/Article.aspx?id=1846.
195. Hussein S, Manthorpe J, Ismail M. Ethnicity at work: the case of British minority workers in the long-term care sector. Equality, Diversity and Inclusion: An International Journal 2014;**33**(2):177-92.
196. Hussein S, Ismail M, Manthorpe J. Changes in turnover and vacancy rates of care workers in England from 2008 to 2010: panel analysis of national workforce data. Health & social care in the community 2015.
197. HSCIC. Personal Social Services: staff of social services departments, England as at September 2014. Leeds: Health and Social Care Information Centre, 2015.
198. Parker J, Doel M, Whitfield J. Does practice learning assist the recruitment and the retention of staff. Research Policy and Planning 2006;**24**(3):179-96.
199. Fenton W. *The size and structure of the adult social care sector and workforce in England, 2011*: Skills for Care, 2011.
200. Skills for Care. The state of the adult social care sector and workforce in England 2015. Leeds: Skills for Care, 2015.
201. Skills for Care. Proposed changes to the Dataset. Secondary Proposed changes to the Dataset 2015. https://www.nmds-sc-online.org.uk/content/view.aspx?id=proposed%20changes.
202. Hussein S, Manthorpe J. Volunteers Supporting Older People in Formal Care Settings in England Personal and Local Factors Influencing Prevalence and Type of Participation. Journal of Applied Gerontology 2014;**33**(8):923-41.
203. Hussein S. Longitudinal Workforce Analysis using Routinely Collected Data: Challenges and Possibilities. London: King's College London, 2012.
204. Hussein S. Migrant workers in long term care: evidence from England on trends, pay and profile. Social care Workforce Periodical 2011(12).
205. Hussein S. *Modelling pay in adult care using linear mixed-effects models*. London: Social Care Workforce Research Unit, Kings College London, 2010.
206. Mindell J, Biddulph JP, Hirani V, et al. Cohort profile: the health survey for England. International journal of epidemiology 2012;**41**(6):1585-93.
207. NatCen. Health Survey for England 2013: User Guide. London: National Centre for Social Research, 2014.

208. Boodhna G, Bridges S, Darton R, et al. (Vol 1): Health, social care and lifestyles. In: Craig R, Mindell J, eds. Health Survey for England 2013. Leeds: Health and Social Care Information Centre, 2014.
209. Scholes S, Coombs N, Pedisic Z, et al. Age-and sex-specific criterion validity of the health survey for England Physical Activity and Sedentary Behavior Assessment Questionnaire as compared with accelerometry. American journal of epidemiology 2014;**179**(12):1493-502.
210. Basterfield L, Adamson AJ, Parkinson KN, et al. Surveillance of physical activity in the UK is flawed: validation of the Health Survey for England Physical Activity Questionnaire. Archives of disease in childhood 2008;**93**(12):1054-58.
211. Tiffin PA, Arnott B, Moore HJ, et al. Modelling the relationship between obesity and mental health in children and adolescents: findings from the Health Survey for England 2007. Child Adolesc Psychiatry Ment Health 2011;**5**(1):31.
212. Boodhna G, Bridges S, Darton R, et al. (Vol 2): Methods and documentation. In: Craig R, Mindell J, eds. Health Survey for England 2013. Leeds: Health and Social Care Information Centre, 2014.
213. Allender S, Foster C, Boxer A. Occupational and non-occupational physical activity and the social determinants of physical activity: results from the Health Survey for England. Journal of physical activity & health 2008;**5**(1):104-16.
214. Nazroo JY, Falaschetti E, Pierce M, et al. Ethnic inequalities in access to and outcomes of healthcare: analysis of the Health Survey for England. Journal of epidemiology and community health 2009;**63**(12):1022-27.
215. Andrew MK. Social capital, health, and care home residence among older adults: A secondary analysis of the Health Survey for England 2000. European Journal of Ageing 2005;**2**(2):137-48.
216. Roth M, Roderick P, Mindell J. Kidney disease and renal function. In: Craig R, Mindell J, eds. Health survey for England, 2010.
217. Thompson J, Wittenberg R, Henderson C, et al. Social care: need for and receipt of help. In: Craig R, Mindell J, eds. (Vol 1) Health, social care and lifestyles. Leeds: Health and Social Care Information Centre, 2014.
218. Neave A. Responses to the Health Survey for England Consultation. Leeds: Health and Social Care Information Centre, 2014.
219. Wheeler BW, Ben-Shlomo Y. Environmental equity, air quality, socioeconomic status, and respiratory health: a linkage analysis of routine data from the Health Survey for England. Journal of epidemiology and community health 2005;**59**(11):948-54.
220. Scholes S, Faulding S, Mindell J. Use of prescribed medicines. In: Craig R, Mindell J, eds. (Vol 1) Health, social care and lifestyles. Leeds: Health and Social Care Information Centre, 2014.
221. Gordon-Dseagu VLZ, Shelton N, Mindell J. Diabetes mellitus and mortality from all-causes, cancer, cardiovascular and respiratory disease: Evidence from the Health Survey for England and Scottish Health Survey cohorts. Journal of Diabetes and its Complications 2014;**28**(6):791-97.
222. Stamatakis E, Ekelund U, Wareham NJ. Temporal trends in physical activity in England: the Health Survey for England 1991 to 2004. Preventive Medicine 2007;**45**(6):416-23.
223. Weston L. Shift Work. In: Craig R, Mindell J, eds. (Vol 1) Health, social care and lifestyles. Leeds: Health and Social Care Information Centre, 2014.
224. Tabassum F, Batty GD. Are Current UK National Institute for Health and Clinical Excellence (NICE) Obesity Risk Guidelines Useful? Cross-Sectional Associations with Cardiovascular Disease Risk Factors in a Large, Representative English Population. PloS one 2013;**8**(7):7.
225. Morris S, Sutton M, Gravelle H. Inequity and inequality in the use of health care in England: an empirical investigation. Social science & medicine 2005;**60**(6):1251-66.
226. Smith N, Malley J. Understanding and addressing underrepresentation in a postal survey of social care users. London: London School of Economics, 2012.

227. Department of Health. The Adult Social Care Outcomes Framework 2015/16. London: Department of Health, 2014.
228. Malley JN, Towers A-M, Netten AP, et al. An assessment of the construct validity of the ASCOT measure of social care-related quality of life with older people. Health Qual Life Outcomes 2012;**10**(21):1477-7525.
229. HSCIC. Personal Social Services Adult Social Care Survey, England 2013-14, Provisional release. Leeds: Health and Social Care Information Centre, 2014.
230. HSCIC. Personal Social Services Adult Social Care Survey, England: Information and guidance for the 2014-15 survey year. Leeds: Health and Social Care Information Centre, 2014.
231. HSCIC. Personal Social Services Adult Social Care Survey, England 2013-14, Final release. Leeds: Health and Social Care Information Centre, 2014.
232. HSCIC. General Practice Extraction Service (GPES): Customer Requirement Summary. Leeds: HSCIC, 2013.
233. NHS England. Care Episode Statistics: Technical Specification of the GP Extract. London: Department of Health, 2013.
234. Boiken E. NHS care.data still leaks like a sinking ship, but ministers set sail regardless. *The Conversation* 2015 June 30th
235. Ramesh R. NHS patient data to be made available for sale to drug and insurance firms. *The Guardian* 2014 January 19th.
236. Donelly L. Hospital records of all NHS patients sold to insurers. *The Telegraph* 2014 February 23rd.
237. Triggle N. Care.data: How did it go so wrong? *BBC* 2014 February 19th.
238. Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care. data ran into trouble. Journal of medical ethics 2015:medethics-2014-102374.
239. Hagger-Johnson GE, Harron K, Goldstein H, et al. THE NHS'S CARE.DATA SCHEME Making a hash of data: what risks to privacy does the NHS's care.data scheme pose? BMJ-British Medical Journal 2014;**348**:1.
240. Renaud-Komiya N. NHS England hits back at highly critical Care.data report. *Health Service Journal* 2015 June 29th
241. CCG BwD. Care.data Update. Secondary Care.data Update 2015. http://www.blackburnwithdarwenccg.nhs.uk/care-data-update/.
242. Statistics: a Data Science for the Twenty-First Century. CHICAS; 2015 June; Lancaster University Medical School.
243. New JP, Bakerly ND, Leather D, et al. Obtaining real-world evidence: the Salford Lung Study. Thorax 2014:thoraxjnl-2014-205259.
244. Greenhalgh T, Stramer K, Bratan T, et al. The devil's in the detail: final report of the independent evaluation of the Summary Care Record and Health Space programmes. 2010. London: University College London, 2010.
245. Harkness EF, Grant L, O'Brien SJ, et al. Using read codes to identify patients with irritable bowel syndrome in general practice: a database study. BMC family practice 2013;**14**(1):183.
246. Davis KJ, New JP, Delderfield MR, et al. Evaluating Feasibility of an EMR-enabled Randomized Clinical Trial in the UK. Secondary Evaluating Feasibility of an EMR-enabled Randomized Clinical Trial in the UK. http://www.idrn.org/documents/events/presentations/primarycaredatabases/Posters/Davis%20K.pdf.
247. Akbarov A, Kontopantelis E, Sperrin M, et al. Primary Care Medication Safety Surveillance with Integrated Primary and Secondary Care Electronic Health Records: A Cross-Sectional Study. Drug safety 2015:1-12.
248. Barnes TRE, Paton C. Role of the Prescribing Observatory for Mental Health. The British Journal of Psychiatry 2012;**201**(6):428-29.
249. Paton C, Barnes TRE, Shingleton-Smith A, et al. Lithium in bipolar and other affective disorders: prescribing practice in the UK. Journal of Psychopharmacology 2010;**24**(12):1739-46.

250. Barnes TRE, Paton C. Improving prescribing practice in psychiatry: The experience of the Prescribing Observatory for Mental Health (POMH-UK). International Review of Psychiatry 2011;**23**(4):328-35.
251. Paton C, Barnes TRE. Undertaking clinical audit, with reference to a Prescribing Observatory for Mental Health audit of lithium monitoring. Psychiatric Bulletin 2014;**38**(3):128-31.
252. Mace S, Taylor D. Reducing the rates of prescribing high-dose antipsychotics and polypharmacy on psychiatric inpatient and intensive care units: results of a 6-year quality improvement programme. Therapeutic advances in psychopharmacology 2014:2045125314558054.
253. Paton C, McIntyre S, Bhatti SF, et al. Medicines reconciliation on admission to inpatient psychiatric care: findings from a UK quality improvement programme. Therapeutic advances in psychopharmacology 2011;**1**(4):101-10.
254. Emilsson L, Lindahl B, Köster M, et al. Review of 103 Swedish healthcare quality registries. Journal of internal medicine 2015;**277**(1):94-136.
255. National Institute for Clinical Excellence. *Principles for best practice in clinical audit*. London: Radcliffe Publishing, 2002.
256. Dixon N. *Ethics and Clinical Audit and Quality Improvement (QI): A Guide for NHS Organisations*. London: HQIP: Healthcare Quality Improvement Partnership, 2009.
257. Dixon N. Research, audit and journal policies. Anaesthesia 2011;**66**(9):847-47.

Appendix 1: List of expert stakeholders

Professor Nick Black	London School of Hygiene and Tropical Medicine
Professor Joanna Chataway	Open University/RAND Europe
Dr José-Luis Fernández	London School of Economics
Colin Flynn	Public Health England
Dr Shereen Hussein	Kings College London
Professor Martin Knapp	London School of Economics
Professor Jan Liliemark	SBU - Swedish Council on Health Technology Assessment
Dr Owen Nicholas	University College London
Dr Miriam O'Hare	Micron Group
Dr Louise Parmenter	Quintiles
Paul Ross	Social Care Institute of Excellence
Professor Tjeerd van Staa	University of Manchester
John Varlow	Health and Social Care Information Centre
Raphael Wittenberg	London School of Economics

*An additional two interviews were carried out with real world data experts based at a regulatory body and a public sector organisation.

Appendix 2: List of sources uncovered

It was also acknowledged that the HSCIC website held a great number of sources that could also be potentially profiled – future exercises could include a more detailed inventory of the HSCIC datasets.

Source name	Type	Disease/Condition /Setting	Population (characteristic)	Geographic	Continuing or Defunct
All Wales Injury Surveillance System	Clinical database	Trauma	All	Wales	Subsumed
Assessment of Stomach and Oesophageal Cancer	Disease registry	Cancer	All	Unknown	Discontinued
British Association of Surgical Oncology—Breast Unit Database	Disease registry	Cancer	All	Unknown	Discontinued
British Isles Network of Congenital Anomaly Registers (BINOCAR)	Disease registry	Congential Abnormalities	Children	Wales, Ireland (NI), 50% of England	Continuing
Carers and Users Expectations of Mental Health Services	Clinical database	Mental Health	All	Unknown	Discontinued
Chronic Obstructive Pulmonary Disease Audits	Clinical audit	Respiratory	All	England and Wales	Continuing
Mothers and Babies: Reducing Risk through Audits and Confidential Enquiries; Centre for Maternal and Child Enquiries (CMACE)	Clinical audit	Maternal health	Mothers and Children	UK	Continuing
East Midlands and South Yorkshire (EMSYCAR)	Disease registry	Congential Abnormalities	Children	Midlands	Continuing
Functional Analysis of Care Environments	Not a primary real world datasource	Mental Health	Unknown	Unknown	Unknown
General Practice Research Database	Clinical database	GP experiences	All	GB	Subsumed
Glasgow Register of Congenital Anomalies	Disease registry	Congential Abnormalities	Children	Scotland	Continuing
Hospital Episode Statistics	Clinical database	Hospital admissions	All	England	Continuing
Intensive Care National Audit and Research Centre—Case Mix Program Database	Clinical database	Trauma	All	UK	Continuing
Manchester Children's Tumour Registry (MCTR); Manchester Children's Cancer registry	Disease registry	Cancer	Children	North England	Continuing
MRC National Survey of Health and Development (1946 cohort)	Survey	Health (general)	Lifecourse; cohort	GB	Continuing
Myocardial Infarction National Audit Programme (MINAP)	Clinical audit	Cardiovascular	Adults	England and Wales	Continuing
National Adult Cardiac Surgery Audit	Clinical audit	Cardiovascular	Adults	England and Wales	Contiuning (2012)
National Drug Treatment Monitoring System	Clinical audit	Substance Abuse	All	England	Continuing
National Pacemaker Database	Surgery/technology register	Cardiovascular	All	UK	Subsumed
National Paediatric Diabetes Audit	Clinical audit	Diabetes	Children	England and Wales	Continuing
National Prospective Monitoring Scheme on HIV	Disease registry	HIV	All	Unknown	Discontinued

National Registry of Childhood Tumours	Disease registry	Cancer	Children	GB	Continuing
National Sentinel Audit of Stroke	Clinical audit	Cardiovascular	All	England, Wales, Northern Ireland	Continuing
North east and North Cumbria (NorCAS)	Disease registry	Congential Abnormalities	Children	North England	Continuing
North of England Collaborative Cerebral Palsy Survey (NECCPS)	Disease registry	Cerebal Palsy	Children	North England	Continuing
North West Hip Arthroplasty Register	Surgery/technology register	General/Technology Assessment	All	North West England	Discontinued
North West Cancer Registry	Disease registry	Cancer	All	North West England	Continuing
Northern Ireland Cancer Registry	Disease registry	Cancer	All	Northern Ireland	Continuing
Northern Region Haematology Register	Disease registry	Cancer	All	North England	Discontinued
Northern Region Young Persons Malignant Disease Registry (NRYPMDR)	Disease registry	Cancer	Children	North England	Continuing
Nosocomial Infection National Surveillance Scheme	Survey	Other Infectious Disease	All	UK	Discontinued
4Child register; Oxford Register of Early Childhood Impairments	Disease registry	Cerebal Palsy	All	South East	Discontinued
Oxfordshire, Berkshire and Buckinghamshire (CAROBB)	Disease registry	Congential Abnormalities	Children	South East	Continuing
Quality Indicators in Diabetes Service	Not a primary real world datasource	Diabetes	Unknown	Unknown	Unknown
Scotland and Newcastle Lymphoma Group Database	Disease registry	Cancer	Unknown	Unknown	Discontinued
Scottish Asthma Management Initiative	Not a primary real world datasource	Respiratory	Unknown	Scotland	Continuing
Scottish Motor Neurone Disease Register	Disease registry	Motor Neurone Disease	All	Scotland	Continuing (2010)
South West Congenital Anomaly Register (SWCAR)	Disease registry	Congential Abnormalities	Children	South West	Continuing
St Mary's Maternity Information System	Clinical database	Maternal health	Women	London	Discontinued
North East Paediatric Diabetes Network (NEPDN) Database	Disease registry	Diabetes	Children	North England	Continuing
Trauma Audit and Research Network	Clinical audit	Trauma	All	UK	Continuing
UK Cystic Fibrosis Database	Disease registry	Genetic disease	All	UK	Continuing
National Diabetes Audit	Clinical audit	Diabetes	All	England and Wales	Continuing
UK Hydrocephalus Shunt Registry	Disease registry	Surgery	All	UK	Continuing
UK Registry for Rare Kidney Diseases	Disease registry	Renal	All	UK	Continuing
UK National Renal Registry	Disease registry	Renal	All	UK	Continuing
UK Register of HIV Seroconverters	Disease registry	HIV	All	UK	Continuing
Wessex Antenatally Detected Anomalies (WANDA)	Disease registry	Congential Abnormalities	Children	South East	Continuing

West Midlands Congenital Anomaly Register (WMCAR)	Disease registry	Congential Abnormalities	Children	Midlands	Continuing
West Midlands Regional Children's Tumour Registry (WMRCTR)	Disease registry	Cancer	Children	Midlands	Continuing
Yorkshire and Humber Congenital Anomaly Register (YHCAR)	Disease registry	Congential Abnormalities	Children	North England	Continuing
Yorkshire Specialist Register of Cancer in Children and Young People (YSRCCYP)	Disease registry	Cancer	Children	North England	Continuing
Tayside Medicines Monitoring Unit (MeMo)	Clinical database	Pharmacoepidemiology	All	Tayside	Continuing
Oxford Record Linkage Study (ORLS)	Clinical database	Hospital admissions	All	Oxfordshire and Berkshire	Discontinued
Human Fertilisation and Embryology Authority Database	Disease registry	Fertility	All	UK	Continuing
Central Cardiac Audit Database (9 individual audits; listed separately)	Disease registry	Cardiovascular	All	UK	Continuing
Yorkshire Register of Diabetes in Children and Young People	Disease registry	Diabetes	Children	Yorkshire	Continuing
Leicestershire Diabetes Register	Disease registry	Diabetes	All	Leicestershire	Discontinued
Breast Implant Register	Surgery/technology register	General/Technology Assessment	Women	Unknown	In development
UKHVR: United Kingdom Heart Valve Register	Surgery/technology register	Cardiovascular	All	UK	Discontinued
National Transplant Database	Donation registry	Organ donation	All	UK	Continuing
NHS Organ Register	Donation registry	Organ donation	All	UK	Continuing
UK Database of Uncertainties about the Effects of Treatments	Not a primary real world datasource	Health (general)	Unknown	Unknown	Continuing
Sample of Anonymised Records (SAR)	Census	Health (general)	All	UK	Continuing
Census - Longitudinal Study (LS)	Census	Health (general)	All	UK	Continuing
British Household Panel Study	Survey	Health (general)	Adults	GB	Subsumed
Understanding Society	Survey	Health (general)	Adults	GB	Continuing
Health Survey for England	Survey	Health (general)	Adults	England	Continuing
ELSA (English Longitudinal Study of Ageing)	Survey	Health (general)/Social Care	Older adults	England	Continuing
Life Opportunities Survey	Survey	Disability	Adults	GB	Continuing
Adult Social Care Survey (ASCS)	Survey	Social Care	Adults	England	Continuing
Count me in census	Census	Mental health	Adults	England and Wales	Discontinued
Children in need census (Looked After Children Stats)	Social care database	Social Care	Children	England	Continuing
National Minimum Data Set for Social Care (NMDS-SC)	Workforce registry	Social Care	Workforce	England	Continuing
The Health Improvement Network (THIN)	Clinical database	GP experiences	All	UK	Continuing
Dr Foster	Not a primary real world datasource	Health (general)	Unknown	Unknown	Continuing

British Thoracic Society National Pleural Procedures Audit	Clinical audit	Respiratory	All	UK	Continuing
British Thoracic Society National Paediatric Bronchiectasis Audit	Clinical audit	Respiratory	Children	UK	Continuing
British Thoracic Society National Paediatric Pneumonia Audit	Clinical audit	Respiratory	Children	UK	Continuing
British Thoracic Society National Adult NIV Audit	Clinical audit	Respiratory	Adults	UK	Continuing
British Thoracic Society National Adult Asthma Audit	Clinical audit	Respiratory	Adults	UK	Continuing
British Thoracic Society National Paediatric Asthma Audit	Clinical audit	Respiratory	Children	UK	Continuing
British Thoracic Society National Adult Bronchiectasis Audit	Clinical audit	Respiratory	Adults	UK	Continuing
British Thoracic Society National Emergency Oxygen Audit	Clinical audit	Respiratory	All	UK	Continuing
British Thoracic Society National Adult Community Acquired Pneumonia Audit	Clinical audit	Respiratory	Adults	UK	Continuing
British Thoracic Society Lung Disease Registry	Disease registry	Respiratory	All	UK	Continuing
British Thoracic Society Difficult Asthma Registry	Disease registry	Respiratory	All	UK	Continuing
The British Society of Urogynaecology (BSUG) Audit Database	Clinical audit	Genitourinary	All	GB	Continuing
National Reporting and Learning System	Incident registry	General/Technology Assessment	All	England	Continuing
National Database for Primary Care Groups and Trusts	GP Population List	Health (general)	All	England	Discontinued
General Household Survey (General Lifestyle Survey)	Survey	Health (general)	Adults	GB	Discontinued
Fourth Morbidity Survey in General Practice (MSGP4)	Survey	Health (general)	All	Unknown	Discontinued
Primary Care Information Services (PRIMIS)	Clinical database	GP experiences	All	England	Continuing
General and Personal Medical Services Data	Workforce registry	Workforce	GPs	England	Continuing (2013)
RCGP RSC National Monitoring Network	Clinical database	GP experiences	All	England and Wales	Continuing
Morbidity, Information Query and Export Syntax (MIQUEST)	Not a primary real world datasource	GP experiences	All	N/A	Continuing
Primary Care Networks: Trent Focus	Not a primary real world datasource	GP experiences	All	Trent	Unknown
Prescribing Analysis and Cost (PACT) data	Pharmacoepidemiological database	Pharmacoepidemiology	All	England	Continuing
Quality Management and Analysis System (QMAS)	Clinical database	GP experiences	All	England	Discontinued
Quality Prevalence and Indicators Database (QPID)	Clinical database	Health (general)	All	England	Subsumed
Qresearch	Clinical database	GP experiences	All	UK	Continuing
UK Biobank	Precision medicine	Biomarkers	Older adults	UK	Continuing
Suicide Information Base Cymru	Mortality registry	Death	All	Wales	Continuing
Welsh Cancer Intelligence & Surveillance Unit (WCISU)	Disease registry	Cancer	All	Wales	Continuing

Welsh Demographic Service (WDS)	Census	Population	All	Wales	Continuing
Primary Care GP dataset	Clinical database	GP experiences	All	Wales	Continuing
National Community Child Health Database (NCCHD)	Screening register	Health (general)	Children	Wales	Continuing
Congenital Anomaly Register and Information Service (CARIS)	Disease registry	Congential Abnormalities	Children	Wales	Continuing
Cervical Screening Wales (CSW)	Screening register	Cancer	Women	Wales	Continuing
Annual District Birth Extract (ADBE)	Birth register	Births	Children	Wales	Continuing
Bowel Screening Wales (BSW)	Screening register	Gastroenterology	All	Wales	Continuing
Scottish Suicide Information Database	Mortality registry	Death	All	Scotland	Continuing
ONS Mortality Dataset	Mortality registry	Death	All	England and Wales	Continuing
Annual District Deaths Extract	Mortality registry	Death	All	Wales	Continuing
Patient Episode Database for Wales	Clinical database	Hospital admissions	All	Wales	Continuing
Outpatient Dataset (OPD)	Clinical database	Hospital admissions	All	Wales	Continuing
Emergency Department Data Set (EDDS)	Clinical database	Trauma	All	Wales	Continuing
Health & Social Care (HSCIC)	National provider of data				
ASCOT (The Anglo Scandinavian Cardiac Outcomes Trial)	Not a primary real world datasource	Cardiovascular	Unknown	Unknown	Unknown
Projecting Older People Population Information (poppi)	Population data	Health (general)	Older adults	England	Continuing
Born in Bradford Study	Survey	Health (general)	Children	Bradford	Continuing
National Child Measurement Programme	Survey	Obesity	Children	England	Continuing
Patients Like Me	Patient reported outcome database	Health (general)	All	UK and International	Continuing
Hospital Pharmacy Audit Index	Pharmacoepidemiological database	Pharmacoepidemiology	All	England	Continuing
Electronic Prescribing Analysis and Cost Tool (ePACT)	Pharmacoepidemiological database	Pharmacoepidemiology	All	England	Continuing
IMS Disease Analyser	Clinical database	GP experiences	All	UK	Continuing
UK Nursing Dataset	Not a primary real world datasource	Nursing	Adults	Unknown	Unknown
General Medical Services (GMS) Data Warehouse	GP Population List	GP experiences	All	Scotland	Continuing
Quality Outcomes Framework	Clinical database/ Disease registry	Health (general)	All	England	Continuing
National Adult Social Care Intelligence Service (NASCIS)	Social care database	Social Care	All	England	Continuing
Transcatether aortic valve intervention (TAVI) Registry	Surgery/technology register	Cardiovascular	All	UK	Continuing
Quality Outcome Framework (QOF) - Dementia Register	Disease registry	Dementia	All	England	Continuing
Prescription Pricing Authority database	Pharmacoepidemiological database	Pharmacoepidemiology	All	England	Continuing
IMS Health databases (Medical Data Index)	Clinical database	Pharmacoepidemiology	All	UK	Continuing

IMS Health databases (MIDAS Prescribing Insights)	Pharmacoepidemiological database	Pharmacoepidemiology	All	UK	Continuing
Yellow card scheme	Disease registry	Adverse reaction/iatrogenic disease	All	UK	Continuing
Smoking Toolkit Study	Survey	Smoking	All	England	Continuing (2014)
Scottish Health Survey	Survey	Health (general)	Adults	Scotland	Continuing
Case Mix Programme (CMP)	Clinical audit	Intensive Care	All	England, Wales, Northern Ireland	Continuing
Scottish Hepatitis-C Virus Clinical database	Disease Registry	Other Infectious Disease	All	Scotland	Continuing
Integrated Household Survey	Survey	Health (general)	All	UK	Continuing
South East London Community Health (SELCoH)	Survey	Mental health	Adults	London	Discontinued
English Adult Psychiatric Morbidity Study (APMS)	Survey	Mental health	Adults	England	Discontinued
West of Scotland Twenty-07 prospective cohort study	Survey	Health (general)	Lifecourse; cohort	Glasgow	Continuing
National Clinical Audit of Falls and Bone Health	Clinical audit	Falls	Older adults	England, Wales, Northern Ireland	Discontinued
The Health and Occupation Reporting network (THOR)	Disease registry	Workforce	All	UK	Continuing
South Yorkshire Cohort health and weight study	Not a primary real world datasource	Obesity	Unknown	Unknown	Unknown
Perineal Assessment and Repair Longitudinal Study (PEARLS)	Not a primary real world datasource	Maternal health	Unknown	Unknown	Unknown
EPIC-Norfolk (European Prospective Investigation of Cancer-Norfolk)	Disease registry	Cancer	All	Norfolk	Continuing
Scottish Morbidity Records	Clinical database	Hospital admissions	All	Scotland	Continuing
Computerised Radiology Information System (CRIS) reports	Not a primary real world datasource	General/Technology Assessment	Unknown	Unknown	Unknown
UK Obstetric Surveillance System	Disease registry	Maternal health	All	UK	Continuing
Prescribing Observatory for Mental Health	Pharmacoepidemiological database	Pharmacoepidemiology (Mental health)	All	UK	Continuing
British Society of Rheumatology Biologics Register (BSRBR)	Pharmacoepidemiological database	Pharmacoepidemiology (Rheumatology)	All	GB	Continuing
National Bowel Cancer Audit	Clinical audit	Cancer	All	England and Wales	Continuing
National Audit of Continence Care for Older People	Clinical audit	Incontinence	Older adults	England, Wales, Northern Ireland	Continuing (2010)
National Cardiac Arrest Audit (NCAA)	Clinical audit	Cardiovascular	All	UK	Continuing
Paediatric Intensive Care Audit Network	Clinical audit	Intensive Care	Children	UK	Continuing
British HIV Association (BHIVA) audit	Clinical audit	HIV	All	GB	Continuing
Scottish Diabetes Survey	Survey	Diabetes	All	Scotland	Continuing

Association of Coloproctology of Great Britain and Ireland (ACPGBI) colorectal cancer database	Clinical audit	Cancer	All	UK	Subsumed
European Clinical Database	Clinical database	Health (general)	All	Europe	Discontinued
Sample of Anonymised Records	Census	Health (general)	All	UK	Continuing
Doctors Independent Network clinical database	Clinical database	GP experiences	All	Unknown	Discontinued
National Child Health Computer System	Clinical database	Health (general)	Children	Wales	Discontinued
Community Mental Health User Survey	Survey	Mental Health	Adults	England	Continuing
Survey of Adult Carers in England (SACE); Carers' Experience Survey (CES)	Survey	Social Care	Adults	England	Continuing
National Child Development Study	Survey	Health (general)	Lifecourse; cohort	GB	Continuing
British Association of Paediatric Surgeons Congenital Anomalies Surveillance System	Disease registry	Congential Abnormalities	Children	UK	Continuing
Hospice Clinical Administration System	Disease registry	Acute care	All	Local	Continuing
Involve to Evolve	Disease registry	Acute care	All	West Midlands	Continuing
Liverpool Orthognathic Database	Surgery/technology register	Surgery	All	Local	Continuing
National InPatient Pain Study	Disease registry	Acute care	All	England, Wales, Northern Ireland	Continuing
North Devon Abdominal Aortic Aneurysm (AAA) Surveillance Register	Disease registry	Acute care	All	Local	Continuing
Podiatric Audit in Surgery and Clinical Outcome Measurement Tool	Disease registry	Podiatry	All	UK	Continuing
Poole Enhanced Recovery Database	Surgery/technology register	Surgery	All	Local	Continuing
Thoracic Surgical Database	Surgery/technology register	Surgery	All	GB	Continuing
UK Fatal Anaphylaxis Register	Disease registry	Acute care	All	England and Scotland	Continuing
UK National HALO Patient Registry	Surgery/technology register	Surgery	All		Continuing
UK Rehabilitation Outcomes Collaborative Database.	Surgery/technology register	Surgery	All	England	Continuing
Databases in Histocompatibility and Immunogenetics	Disease registry	Blood disorder	All	GB	Continuing
Databases in Red Cell Immunohematology	Disease registry	Blood disorder	All	GB	Continuing
National Haemoglobinopathy Registry	Disease registry	Blood disorder	All	GB	Continuing
Association of Breast Surgery Breast Screening Audit Database	Screening register	Cancer	All	GB	Continuing
BAUS Cancer Registry (BCR) Complex Operations Audit	Surgery/technology register	Cancer	All	GB	Continuing

Breast Cancer Clinical Outcome Measures Project	Disease registry	Cancer	All	GB	Continuing
British Association of Urological Surgeons Cancer Registry	Disease registry	Cancer	All	GB	Discontinued
Eastern Cancer Registration and Information Centre	Disease registry	Cancer	All	GB	Continuing
Family History of Bowel Cancer Clinic Database	Disease registry	Cancer	All	GB	Continuing
KC65	Screening register	Cancer	All	Local	Continuing
Liverpool Head and Neck Cancer	Disease registry	Cancer	All	Local	Continuing
National Botulinum Toxin Therapy Audit	Disease registry	Cancer	All	GB	Continuing
Sloane Project	Disease registry	Cancer	All	GB	Continuing
Thames Cancer Registry	Disease registry	Cancer	All	South East	Continuing
Upper Urinary Tract Transitional Cell Carcinoma Audit	Disease registry	Cancer	All	GB	Continuing
Welsh Cancer Intelligence and Surveillance Unit	Disease registry	Cancer	All	Wales	Continuing
National Cancer Intelligence Network (various data)	Disease registry	Cancer	All	GB	Continuing
Lipoprotein Apheresis Register	Disease registry	Cardiovascular	All	GB	Continuing
Sudden Arrhythmic Death Syndrome Database	Disease registry	Cardiovascular	All	GB	Continuing
International Gastrointestinal Neuromuscular Disease Database	Disease registry	Gastroenterology	All	England	Continuing
Alström Syndrome UK Clinical Research Database	Disease registry	Genetic disease	All	GB	Continuing
Clinical and Laboratory Online Patient- and Research Database for Primary Immunodeficiencies	Disease registry	Immunodeficiencies	All	GB	Continuing
Database of Alkaptonuria Patients	Disease registry	Genetic disease	All	GB	Continuing
Diabetes Specialist Workforce Audits	Workforce registry	Workforce	Workforce	GB	Continuing
DiabetesE	Service audit	Workforce	Workforce	England	Continuing
Early Rheumatoid Arthritis Study	Disease registry	Rheumatology	All	Local	Continuing
East Kent Carpal Tunnel Syndrome Database	Disease registry	Other	All	South East	Continuing
East Kent Movement Disorders Database	Disease registry	Other	All	South East	Continuing
Hepatology Database	Disease registry	Liver disease	All	Local	Continuing
HIV and AIDS New Diagnoses Database	Disease registry	HIV	All	GB	Continuing
IBD database (Infoflex)	Disease registry	Gastroenterology	All	Local	Continuing
IBD Registry (Dendrite)	Disease registry	Gastroenterology	All	England	Continuing
Inflammatory Bowel Disease Database	Disease registry	Gastroenterology	All	North East	Continuing
National Congenital Rubella Surveillance Programme	Disease registry	Congential Abnormalities	All	UK	Continuing
National Haemophilia Database	Disease registry	Blood disorder	All	UK	Continuing

National New Patient Registration of Patients with Bisphosphonate Related Osteonecrotic Jaw in England, Wales, Scotland & N Ireland	Disease registry	Other	All	UK	Discontinued
Sarcoidosis Registry	Disease registry	Immunodeficiencies	All	UK	Continuing
Survey of Prevalent HIV Infections Diagnosed	Disease registry	HIV	All	UK	Continuing
UK Collaborative HIV Cohort Study	Survey	HIV	All	London and Scotland	Continuing
UK HIV Drug Resistance Database	Pharmacoepidemiological database	Pharmacoepidemiology (HIV)	All	UK	Continuing
UK National Neuromuscular Database	Disease registry	Other	All	GB	Continuing
Unique Database	Disease registry	Genetic disease	All	GB	Continuing
National Ophthalmology Database	Disease registry	Opthalmology	All	GB	Continuing
Non Arthroplasty Hip Register	Disease registry	Surgery	All	UK	Continuing
Temporomandibular joint prosthesis national registration	Disease registry	Other	All	UK	Continuing
All Wales Perinatal Survey	Mortality registry	Death	Children	Wales	Continuing
CRANE Database	Disease registry	Congential Abnormalities	Children	England, Wales, Northern Ireland	Continuing
National Study of HIV in Pregnancy and Childhood	Disease registry	HIV	Children	UK	Continuing
Northern Survey of Diabetes in Pregnancy	Disease registry	Diabetes	Children	North East England and North Cumbria	Continuing
Northern Survey of Twins and Multiple Pregnancy	Mortality registry	Death	Children	North East England and North Cumbria	Continuing
Perinatal Mortality and Morbidity Survey	Mortality registry	Death	Children	North East England and North Cumbria	Continuing
National Neonatal Research Database	Clinical database	Health (general)	Children	England and Wales	Continuing
Hip Fracture Perioperative Network	Clinical database	Bone, Joint, Muscle	All	UK	Continuing
International Burn Injury Database	Clinical database	Trauma	All	England and Wales	Continuing
Clinical Database for Cluster Headache	Disease registry	Other	All	UK	Continuing
British Society of Interventional Radiology (BSIR) Inferior Vena Cava (IVC) Filter Registry	Surgery/technology register	Surgery	All	UK	Continuing
Bowel cancer audit	Clinical audit	Cancer	All	GB	Continuing
Cardiac arrhythmia audit	Clinical audit	Cardiovascular	All	GB	Continuing
Chronic Kidney Disease in primary care audit	Clinical audit	Renal	All	GB	Continuing
Congenital heart disease (Paediatric cardiac surgery)	Clinical audit	Cardiovascular	Children	GB	Continuing

Coronary angioplasty / purcutaneous coronary interventions	Clinical audit	Cardiovascular	All	GB	Continuing
Diabetes Audit (Adult)	Clinical audit	Diabetes	Adults	GB	Continuing
Diabetes Audit (Paediatric)	Clinical audit	Diabetes	Children	GB	Continuing
Falls and Fragility Fractures Audit Programme (includes the hip fracture database)	Clinical audit	Falls	All	GB	Continuing
Head and Neck Oncology Audit	Clinical audit	Cancer	All	GB	Continuing
Inflammatory Bowel Disease Audit	Clinical audit	Gastroenterology	All	GB	Continuing
National Emergency Laparotomy Audit	Clinical audit	Surgery	All	GB	Continuing
National Joint Registry	Clinical audit	Bone, Joint, Muscle	All	GB	Continuing
National Vascular Registry	Clinical audit	Cardiovascular		GB	Continuing
Neonatal intensive and special care Audit	Clinical audit	Trauma	Children	GB	Continuing
Oesophago-gastric cancer Audit	Clinical audit	Cancer		GB	Continuing
Ophthalmology Audit	Clinical audit	Opthalmology	Men	GB	Continuing
Prostate cancer	Clinical audit	Cancer	All	GB	Continuing
Rheumatoid and early inflammatory arthritis Audit	Clinical audit	Rheumatology	All	GB	Continuing
End of Life Care Audit	Clinical audit	Palliative Care	All	GB	Continuing
Cohort for Skeletal Health in Bristol and Avon (COSHIBA)	Cohort study	Musculoskeletal	Women	SW England	Continuing
UK 10K Rare Genetic Variants in Health and Disease	Research database	Precision medicine	Unknown	Various	Continuing
Twins UK Cohort	Cohort study	Precision medicine	Adults	UK	Continuing
National Safety Thermometer	Survey	Patient safety	Adults	UK	Continuing
Mental Health & Learning Disabilites Data Set (MHLDDS)	Clincal Audit	Mental Health	Adults	England	Continuing
Learning Disability Census	Clincal Audit	Learning Disability	Adults	England	Continuing
Maternity Service Dataset	Clinical Audit	Maternal health	Women and child	England	Continuing
Child Dental Health Survey	Survey	Dental care	Children	England, Wales, Northern Ireland	Continuing
Patient Experience of Diabetes (PEDS)	Survey	Diabetes	All	England and Wales	Continuing
Critical Care Minimum Data Set (CCMDS)	Clinical audit	Critical/emergency care	Adults	England	Continuing
National InPatient Diabetes Audit (NaDIA)	Clinical audit	Diabetes	All	England	Continuing
NHS Dental Statistics for England	Service audit	Dental care	Adults	England	Continuing
The National Audit of Pulmonary Hypertension (NAPH)	Clinical audit	Respiratory	Adults	England, Scotland, Wales, Northern Ireland, the Channel Islands,	Continuing

				Gibraltar and the Isle of Man	
The Improving Access to Psychological Therapies (IAPT) Dataset	Service audit	Mental Health	Adults	England	Continuing
The Adult Psychiatric Morbidity Survey (APMS)	Survey	Mental Health	Adults	England	Continuing
GP Practice Prescribing Presentation-level Data	Clinical database	Medical prescriptions	All	England	Continuing
Cardiovascular disease profile/ National Cardiovascular Intelligence Network (NCVIN)	Disease registry	Cardiovascular	All	England	Continuing
Diabetes Community Health Profile	Disease registry	Diabetes	Adults	England	Continuing
Interactive Health Atlas of Lung Conditions in England (INHALE)	Disease registry	Respiratory	All	England	Continuing
General Lifestyles Survey	Survey	Public Health	Adults	GB	Discontinued
National Centre for Smoking Cessation and Training Statistics	Clinical database	Smoking	All	England	Continuing

Appendix 3: Semi-structured interview schedules

Social care example

- Thank you for time
- Introduce – Researcher; EPPI-Centre
- Working on a number of projects looking at the way evidence is used on decision-making, including this one for NICE
- Overall we're seeing the number of RW datasources expand but also the analytical capability is also
- We know that there are many sources – but they're also fragmented and of varying quality
- But it is of value in assessing practice and trends in real world settings – and in many ways the diversity may be highly positive; – NICE would like to use more in its work
- NICE want to use the data in the following ways:

(f) **Research the effectiveness of interventions or practice** in real-world (UK) settings (e.g. through monitoring outcomes or proxy outcomes). E.g. NICE has a choice between two drugs for migraine and wants to know about adverse complications in routine settings.

(g) **Audit the implementation of guidance.** For example, to assess the equity of implementation across different groups (including socioeconomic, geographic, demographic and groups differentiated by different diseases/health conditions)

(h) **Provide information on resource use** and evaluate the potential impact of guidance.

(i) **Provide epidemiologic information.** For example prevalence/incidence of diseases, natural history, co-morbidities and information on current practice.

(j) **Provide information on current practice to inform the development of NICE quality standards** – e.g. what works best, so for example minimum staffing levels

- **Interview** will take approximately 45-1 hr but can be shorter if have other commitments
- **Being recorded** – check if okay
- **Check if okay** with information and consent sheet
- **Going** to be talking about RW data – and we'll be asking about definition – but for now referring to it as data collected outside RCT
- **Most questions** open about experiences of using different sources, in what way, and thinking about how assessing strengths and weaknesses

Questions to ask:

Appendix 3: Semi-structured interview schedules

	Construct	Question	Probes (if needed)
1	Opening	Can you tell me a little about your experience or interest in using RW data?	Maybe a little bit about any studies using RW data that you've been involved with in the past couple of years
2	Suitability of real-world data for social care	RW data tends to refer to routinely collected administrative data that is collected outside laboratory or experimental conditions. A fundamental principle of real-world data is that no 'treatment'/'technology' or in this case social care package would be changed on account of the collection of the real-world data itself. Is this a workable principle for social care real world data?	Should a different definition of RW data be considered for social care RW data as opposed to data from other fields?
3	Types of real world data coverage (i)	RW data covers a broad taxonomy of different sources and forms. Some of the different forms include: - Administrative data - Routinely collected data (e.g. based on LA returns) - Needs/care registries - Workforce registries How well represented are the different forms in terms of social care real world data sources? Is social care RW data more reliant on survey based methods?	Are there some forms of real-world data that are better represented than others in terms of data from social care settings?
4	Types of real world data coverage (ii)	Are there some groups in receipt of social care that are better or worse represented than others in terms of RW data? For example are	Are there some population groups in terms of social care where more should be done to collect data?

		data lacking on the breadth of older people in receipt of social care, but are data on adults with learning difficulties better represented?	
5	Types of real world data coverage (iii)	Are there some **settings** where there is greater representation in terms of RW settings but some setting where there is underrepresentation? For example are there problems with getting private sector data?	Given that many social care settings actually represent private provider settings, are there initiatives out there to encourage collection from these?
6	Types of real world data – suitability for social care	Are there some forms of RW data that are better suited for collection in social care settings? For example are surveys a more appropriate tool in social care settings?	
7	Experience – go to sources for different purposes	I want to move on to some questions about your own experience of using RW data. Firstly would you be able to let me know what might be your first port of call in terms of RW data when it comes to assessing practice for a particular group – e.g. older people. So these might be data which include depth around the characteristics of older people, the service they receive, and their outcomes (almost like a gold standard)? How would you appraise the quality of this data?	Aggregate data
8	Experience – go to sources for different purposes	Suppose NICE were to recommend a particular care pathway (e.g. for older people) and you were interested in examining the extent to which care providers were adhering to this recommendations – where	Best practice; Safeguarding adults database

		might be your first port of call in terms of RW data when it comes to assessing this type of problem?	
		How would you appraise the quality of this data?	
9	Experience – go to sources for different purposes	Suppose NICE were interested in examining resource use – e.g. staff time and costs (e.g. for older people) – where might be your first port of call in terms of RW data when it comes to assessing this type of problem?	
		How would you appraise the quality of this data?	
10	Experience – go to sources for different purposes	Suppose NICE were interested in examining the prevalence of care needs – e.g. through ADLs– where might be your first port of call in terms of RW data when it comes to assessing this type of problem?	
		How would you appraise the quality of this data?	
11	Challenges	How might you summarise the main challenges of working with RW based on your experiences?	Particular challenges around real world data
12	Ways of overcoming challenges	Which new sources of data do you see helping to overcome some of the challenges you mentioned earlier [repeat challenge back]	
13	Closing	Is there anything we haven't covered which you think we should consider in our study?	

Appendix 4: Current real-world data usage by NICE

Grey cells indicate where the data are not in use; green cells are those that do not constitute real-world data or are incorrectly specified

NICE team	Organ-isation	Title of dataset or database	Licence/subscription holder arrangements	Costs (for subscription/usage)	What the data is used for	Limitations/Comments
Costing and Commissioning, Health & Social Care	HSCIC	Primary Care - IMS disease analyser. Secondary Care (HES) - Admitted care, outpatients, maternity, A&E, Adult critical care.	Licence	£xk per user; 6/7 users	Costing of guidance and QS; Impact & evaulation	
Costing and Commissioning, Health & Social Care	NHS Prescription Services	ePACT - Prescribing data - Primary & community. Also includes non NHS pharmacists located on NHS premises	Agreement	None	Drug prescribing information	
Costing and Commissioning, Health & Social Care	IMS Health	HPIA - hospital dispensed prescribing data	Agreement	None	Drug prescribing information	
Impact & Evaluation	NHS Business Services Authority	Electronic prescribing analysis and cost tool system (EPACT)	The I&E data analysts can access ePACT data via a network connection.	None	To inform medicines metrics work, monitoring the uptake of NICE guidance. Ad hoc enquiries from other teams within NICE.	ePACT data do not link to patient information.

Appendix 4: Current real-world data usage by NICE

Impact & Evaluation	IMS (via HSCIC)	IMS HEALTH Hospital Pharmacy Audit Index (IMS HPAI)	The I&E data analysts request data from the Health and Social Care Information Centre (HSCIC).	None	To inform medicines metrics work, monitoring the uptake of NICE guidance. Ad hoc enquiries from other teams within NICE.	The data do not link to patient information.
Impact & Evaluation	Access via HSCIC	The Health Improvement Network (THIN) database	The I&E data analysts request data from the Health and Social Care Information Centre (HSCIC).	None	To inform medicines metrics work, monitoring the uptake of NICE guidance. Ad hoc enquiries from other teams within NICE.	Vision software is patient management and not medical research; data reflects events deemed to be relevant to the patient's care only.
Impact & Evaluation	Access via HSCIC	IMS Disease analyser	The I&E data analysts request data from the Health and Social Care Information Centre (HSCIC).	None	To inform medicines metrics work, monitoring the uptake of NICE guidance. Ad hoc enquiries from other teams within NICE.	
Impact & Evaluation	HSCIC	Quality Outcomes Framework	QOF results are pubically available as such the analysts have access to the data via HSCIC website.	None	To inform medicines metrics work, monitoring the uptake of NICE guidance. Ad hoc enquiries from other teams within NICE.	
Impact & Evaluation	HSCIC	Hospital Episode Statistics (HES)	The data analysts have been trained to query the database using SAS Enterprise guide, to produce bespoke queries and NICE has purchased licenses to access the database.	The cost of the primary license holder/user for a new organisation is £xxxx. All additional license holders/users within the same organisation cost £xxxx per license	To inform medicines metrics work, monitoring the uptake of NICE guidance. Ad hoc enquiries from other teams within NICE.	Also used by Interventional procedures

Appendix 4: Current real-world data usage by NICE

Safe Staffing (and contractors)	HSCIC	Health Episode Statistics (HES)	Contractor request via official application for data on website	Variable (approximately £xxxx)	To inform safe staffing evidence review, economic analysis and modelling. Estimating relationship between staffing numbers, and skill mix to patient outcomes using multilevel, multivariate regression analysis.	Data requires careful cleaning and filtering. NICE already have access as members of the I&E team are trained to extract data from HES and we have direct access to it. Any team can request data from I&E within NICE.
Impact & Evaluation		ONS		None	To inform medicines metrics work, monitoring the uptake of NICE guidance. Ad hoc enquiries from other teams within NICE. The analysts use ONS data for population and epidemiology data.	Also used by Interventional procedures
Impact & Evaluation		National audits		None		
Impact & Evaluation		Uptake collection / ERNIE		None	To inform medicines metrics work, monitoring the uptake of NICE guidance. Ad hoc enquiries from other teams within NICE.	Summary of other data sources.

Appendix 4: Current real-world data usage by NICE

Safe Staffing (and contractors), Costing and Commissioning	HSCIC	Workforce data / ESR	NICE: analysts request data from the Health and Social Care Information Centre (HSCIC). Contractors via official application for data on website	None	To inform safe staffing economic analysis and modelling. Informs for costing commentary on safe staffing guidelines. Contains baseline registered nurses, healthcare assistants, and other professional baseline numbers per organisation.	Speciality of ward-type information at a very high-level.
Safe Staffing (and contractors)		UK Nursing database (Keith Hurst dataset)	Direct request via email	Variable (approximately £xxxx)	Workforce (nursing) data and outcomes. To inform safe staffing evidence review, economic analysis and modelling. Estimating relationship between staffing numbers, and skill mix to patient outcomes using multivariate regression analysis.	Data collected at irregular intervals at some hospitals. Requires substantial cleaning before statistical analysis. Quality of data highly variable.
Social care	HSCIC	Social care activity, social care expenditure	Accessed from HSCIC website direct	None	To inform the writing of scopes. For example HSCIC has a national data collection that reports information on the numbers of people accessing home care by local authority and expenditure on home care by local authority.	Nationally reported datasets do not always provide data in the detail that can be obtained from bespoke data requests from the HSCIC. Lacking in particular is details on people purchasing their own care

Appendix 4: Current real-world data usage by NICE

Guidance producing /supporting teams	NHSE	Systemic Anti Cancer Therapy (SACT) dataset	Under discussion	TBD	To inform on the uptake of cancer medicines.	We are currently in discussion with NHSE, who manage the SACT dataset, about how we can access this and what we can use it for.
Internal Clinical Guidelines (CCP)	HSCIC	The Health Improvement Network (THIN) database	Directly request data from the Health and Social Care Information Centre (HSCIC).	None	Health economic modelling (e.g. defining baseline characteristics of population of interest)	No ethnicity data
Internal Clinical Guidelines (CCP)	HSCIC	Hospital Episode Statistics (HES)	see I&E arrangements	see I&E arrangements	Health economic modelling (e.g. estimating resource use associated with particular types of hospital admission)	
Internal Clinical Guidelines (CCP)	HSCIC	Health Survey for England	n/a	n/a	Health economic modelling (can be good for quality of life data; have used to plug the ethnicity hole left by THIN)	
Internal Clinical Guidelines (CCP)	DH	NHS reference costs	n/a	n/a	Health economic modelling (mainly costs, obviously, but also some useful data on frequency of hospital contacts and duration of admissions)	

Appendix 4: Current real-world data usage by NICE

Internal Clinical Guidelines (CCP)	NHS Prescription Services	Prescription Cost Analysis (PCA) Data	n/a	n/a	Health economic modelling and other cost analyses	
Internal Clinical Guidelines (CCP)	NHS Prescription Services	NHS drug tariff	n/a	n/a	Drug costs for health economic modelling and other cost analyses	Not all medications covered
Internal Clinical Guidelines (CCP)	eMC	Dictionary of Medicines and Devices	n/a	n/a	Drug costs for health economic modelling and other cost analyses (especially in instances where products are not listed in NHS drug tariff)	
Internal Clinical Guidelines (CCP)	ONS	National Life Tables	n/a	n/a	Health economic modelling	
Internal Clinical Guidelines (CCP)	ONS	Cancer Registration Statistics	n/a	n/a	Health economic modelling	
National Collaborating Centre - Mental Health	HSCIC	Health Episode Statistics (HES)	none at the moment	N/A	to inform guideline models	
National Collaborating Centre - Mental Health	NHS	NHS reference costs	freely available	None	Health economic modelling	rough coding of MH services on some occasions (not always disorder-specific)
National Clinical Guideline Centre	DH	NHS reference costs	Data are freely available on website	None	Health economic modelling	https://www.gov.uk/government/publications/nhs-reference-costs-2013-to-2014

National Clinical Guideline Centre	HSCIC	Health Episode Statistics (HES)	Data are freely available on website	None	Health economic modelling	
National Clinical Guideline Centre	PSSRU	Unit costs of health and social care	Data are freely available on website	None	Health economic modelling	http://www.pssru.ac.uk/project-pages/unit-costs/
National Clinical Guideline Centre	NHSBSA	Drug tariff	Data are freely available on website	None	Health economic modelling	http://www.nhsbsa.nhs.uk/PrescriptionServices/4940.aspx
National Clinical Guideline Centre	NICE	BNF	Data are freely available on website	None	Health economic modelling	
National Clinical Guideline Centre	NHSBSA	Prescription Cost Analysis (PCA) Data	Data are freely available on website	None	Health economic modelling	http://www.nhsbsa.nhs.uk/PrescriptionServices/3494.aspx
National Clinical Guideline Centre	DH CMU	EMIT	Data are freely available on website	None	Health economic modelling	https://www.gov.uk/government/publications/drugs-and-pharmaceutical-electronic-market-information-emit
National Clinical Guideline Centre	OECD	Purchasing power parities	Data are freely available on website	None	to convert the results of overseas economic evaluations to £	http://stats.oecd.org/Index.aspx?datasetcode=SNA_TABLE4
National Clinical Guideline Centre	ONS	National Life Tables & mortality data	Data are freely available on website	None	Health economic modelling	http://www.ons.gov.uk/ons/taxonomy/index.html?nscl=Interim+Life+Tables#tab-data-tables
NCC-Cancer	Quality Health	National Cancer Patient Experience Survey	Data supplied via e-mail	None	People's experience of cancer diagnosis and	

					treatment- according to ICD-10 code.	

The Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) is part of the Social Science Research Unit (SSRU), UCL Institute of Education, University College London.

The EPPI-Centre was established in 1993 to address the need for a systematic approach to the organisation and review of evidence-based work on social interventions. The work and publications of the Centre engage health and education policy makers, practitioners and service users in discussions about how researchers can make their work more relevant and how to use research findings.

Founded in 1990, the Social Science Research Unit (SSRU) is based at the UCL Institute of Education, University College London. Our mission is to engage in and otherwise promote rigorous, ethical and participative social research as well as to support evidence-informed public policy and practice across a range of domains including education, health and welfare, guided by a concern for human rights, social justice and the development of human potential.

The views expressed in this work are those of the authors and do not necessarily reflect the views of the EPPI-Centre or the funder. All errors and omissions remain those of the authors.

First produced in 2016 by:

Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre)
Social Science Research Unit
UCL Institute of Education, University College London
18 Woburn Square
London WC1H 0NR

Tel: +44 (0)20 7612 6391

http://eppi.ioe.ac.uk/
http://www.ioe.ac.uk/ssru/

ISBN: 978-1-907345-93-7

This document is available in a range of accessible formats including large print.

Please contact the UCL Institute of Education for assistance:
telephone: +44 (0)20 7947 9556 email: info@ioe.ac.uk