University of BRISTOL

Economic and Social Research Council

Understanding Population Data for Inclusive Longitudinal Research

Andy Boyd

October 2021

Correspondence:
Andy Boyd
University of Bristol
a.w.boyd@bristol.ac.uk

Citation of the Report
Boyd A. (2021). Understanding Population Data for Inclusive Longitudinal Research. Bristol, UK: University of Bristol.

Acknowledgements

This report is dedicated to Harvey Goldstein who died from Covid-19 at the start of the pandemic in March 2020. Harvey was a friend and mentor: a source of endless support and guidance. His friendship and contributions to my thinking on the subjects of data science and longitudinal research are very much missed.

This scoping review was commissioned and funded by the Economic and Social Research Council (ESRC Ref: ES/S016732/1) as part of their Population Data Laboratory programme of work.

A very large number of individuals and organisations have contributed to this study. Many thanks to all those working for longitudinal studies, in the NHS and Office for National Statistics and other government agencies who contributed evidence to this report. I would particularly like to thank Emma Gordon and Paul Jackson (ADRUK), David Ford (SAIL Databank and ADRC-Wales), Garry Coleman (NHS Digital), Katie Harron and Harvey Goldstein (UCL & University of Bristol), Miranda Mourby (University of Oxford) and John Macleod (University of Bristol) for the multiple times I turned to you for input. Your help, guidance and insights were very much appreciated. I also thank the Economic and Social Research Council which made my secondment to undertake this study so enjoyable and insightful. Particular thanks to Rebecca Perring for all her valuable contributions, support and insights and for the guidance and support of the wider team of Bridget Taylor, Catherine Bromley, Robert Wilson and Thomas Graham.

Thank you to the ALSPAC cohort study and University of Bristol for agreeing to support my secondment which enabled this study. Much of the thinking leading up to this study was based on work at ALSPAC (MRC/WT Ref: 217065/Z/19/Z), particularly the PEARL project to enhance linkages in ALSPAC (WT Ref: 086118 & MRC Ref: MC_PC_17210) and the ERICA project developing geospatial linkages (NERC Ref: R8/H12/83/NE/P01830/1); and, also my work as Data Linkage Theme Lead at the CLOSER longitudinal research consortium (ESRC Ref: ES/K000357/1). Whilst the majority of the report was completed prior to the COVID-19 pandemic, the report has been updated to include emerging developments including the UK Longitudinal Linkage Collaboration, which is a programme within the Longitudinal Health & Wellbeing National Core Study (UKRI Ref: MC_PC_20059), and the work of the Data & Connectivity National Core Study (UKRI Ref: MC_PC_20058).

A particular thanks is made to the 2-3 million UK residents who have given up so much of their time to longitudinal research with the aim of contributing to the public good. I was very mindful of the trust you place in our community while conducting this study; and the expectation you place on us to protect your rights whilst maximising the research value of the data you donate.

Andy Boyd
Population Health Sciences
University of Bristol

26th October 2021.

Executive Summary

Background: The 2018 report to the Economic and Social Research Council (ESRC) by the independent Longitudinal Studies Strategic Review (LS Review) panel identified a risk that UK longitudinal population studies (LPS) lacked 'representativeness' and that harder to reach vulnerable and marginalised populations may not be well served by longitudinal research and are missing the benefits this brings. Since the LS Review, the Covid-19 pandemic and governmental and societal emphasis on inequalities, regional differences in opportunities and social justice have reinforced the importance of equity in longitudinal research. The LS Review suggests this may be addressed by better use of population data: particularly by developing a whole population 'Administrative Data Spine' (ADS) register for sampling, recruitment, coverage assessment and follow-up.

Scoping study remit: The ESRC commissioned this study to gather interdisciplinary evidence on how population data can help ensure inclusive longitudinal research and to identify:

- The population coverage of UK LPS, and missing populations, to help ensure that any new LPS is inclusive and, where necessary, representative of the UK population.
- The population data sets that are in operation in the UK and to understand what inclusive infrastructure and methods look like and the ethico-legal basis for these;

Key findings and recommendations

1. **An identifiable whole UK population 'Administrative Data Spine' to support research is legally and technically feasible but not proportionate or acceptable.**

It is recommended that the ESRC should not pursue the ADS option at present.

The scoping study has identified that the proposed ADS model is technically feasible but neither proportionate (in terms of costs and impact on personal privacy) nor acceptable (there is substantial evidence that there is no political will for this and that public aversion to this way of working could render it unviable). Yet, the Covid-19 pandemic has led to new ways of working with population data, and there is benefit in considering how functions of the ADS could be included in any considerations of data needs to support public service delivery, statistics and research. This study has identified that the most suitable alternative datasets for sampling and recruiting LPS participants are presently the birth register and national NHS Patient Registers.

2. **Sampling selection and recruitment could be more efficient where informed by individual level population data**

It is recommended that privacy preserving protocols are considered to mitigate perceived privacy risks during sample selection and recruitment where access to identifiable data for opt-out recruitment approaches cannot be secured.

Sampling and recruitment are often informed by area rather than individual level socio-economic indicators due to barriers in accessing individual data for opt-out recruitment approaches: privacy preserving protocols may address barriers and enable dynamic recruitment to help realise substantial recruitment efficiencies, potential improvements in sample heterogeneity and targeting of resources to harder to reach groups.

3. **There is an ethical and legal obligation to be inclusive in LPS research**

It was identified that there is a legal duty for those developing UK LPS strategic thinking, to consider inclusion, fairness and equality at the level of the longitudinal community.

It is recommended that the ESRC consider non-statutory options for including longitudinal research and data science concepts in the UK national curriculum to improve awareness and to help sustain the 'social licence' for using population data.

The study found that LPS have a social and ethical obligation to conduct high-quality research that is inclusive of vulnerable and disadvantaged groups. It also, for the first time, recognised the duty that those developing UK LPS strategy have under Equalities legislation. The public should be empowered to understand and help shape LPS strategies; secondary-school teaching of LPS methods, benefits and safeguards could help foster a willingness to participate.

4. **Long term follow-up of participants through linkage infrastructure**

It is recommended that the LPS community develop an interdisciplinary centralised Trusted Research Environment for linking study and population data and that this is supported by LPS funders: the UK Longitudinal Linkage Collaboration (UK LLC) for Covid-19 research forms a model for this.

The study highlights that different population groups are differentially harder to recruit and retain and that population data can be used to assess and avoid some bias resulting from this. However, barriers to data access and changing governance frameworks have resulted in uneven implementation of linkage in LPS. For sustainable linkage follow-up, a community consortium model could be acceptable to studies, participants and other key stakeholders with appropriate safeguards. The UK LLC has brought together data from over 20 interdisciplinary UK studies and is systematically linking these to Covid-19 relevant data under a single governance framework which accommodates study needs. This model draws on learning from this study and this (or similar) should be sustained and generalised to wider use purposes with participant involvement and in a transparent and well-defined manner.

5. **There is a need to assess diversity and inclusion across LPS at a community level**

It is recommended that sample diversity and follow-up are systematically assessed to consider how the sum total of LPS is inclusive of the UK population.

The study identified good practice and commitment from LPS to inclusive research. Yet, it also found suggestions that vulnerable and marginalised groups are disproportionately missing from LPS and that successful engagement strategies for harder to reach groups are based on long-term trust relationships. LPS should develop and implement evidence-based 'Inclusion plans' with input from participants and impacted groups; LPS funders should resource these and monitor delivery using relevant metrics. LPS must publish evidence of sample composition so LPS funders can fulfil their legal duty. The UK LLC (or similar) may provide a systematic linkage informed means to achieve this.

Next steps: The LPS community should work with representatives of vulnerable groups and more widely with participants and the public to explore how Covid-19 ways of working can be sustained with the aim of developing public goods through more inclusive research. LPS funders should evaluate their funding schemes and infrastructure management processes to help ensure more Equitable longitudinal research.

Contents

Chapter 1: Introduction, background and methods

Introduction

1.1 High-quality evidence is needed to support the understanding of health and social phenomena and to inform the development of government policy and service provision. Longitudinal Population Studies (LPS)[1] help provide such evidence by collecting the diverse longitudinal data needed to inform assessments of how biological, health, socio-economic and environmental factors interact to influence a broad range of outcomes.

1.2 This report presents the findings of a scoping study conducted following the publication of The Economic and Social Research Council's (ESRC) Longitudinal Studies Strategic Review (LS Review). The purpose of the scoping study is to examine whether UK longitudinal studies' samples are inclusive of different population groups and those with different health and social circumstances; and whether this means some groups - potentially the most vulnerable and marginalised - are not fully served by longitudinal research and the benefits it can bring. The scoping study specifically considers the role of population data in helping to investigate and address this challenge.

1.3 The scoping study was conducted prior to the Sars-cov-2 (Covid-19) pandemic. The pandemic has generated a pressing need for data and evidence as to the impacts of the Covid-19 virus and insights into how Covid-19 mitigations are impacting on wider aspects of health and wellbeing. This requirement - alongside wider appreciation of health and social inequalities - is resulting in new approaches to the flow of population data for research with new research possibilities. The report has been updated to reflect this, although no new evidence gathering has taken place.

Background

The Economic and Social Research Council Longitudinal Studies Strategic Review

1.4 In 2017 the ESRC published its LS Review. The aim of the review was:

"to provide an evidence-based and challenge-led assessment of the future social and interdisciplinary scientific and policy-relevant needs for data to address the types of research questions for which longitudinal data has typically been used (or could be used), and the value of the life-course evidence from our longitudinal studies in comparison with other sources of evidence." (Davis-Kean *et al*, 2017; p. 48).

1.5 The LS Review emphasised the value of LPS to those considering research and policy questions and reflected that the UK has a strategic advantage in that its world leading investment in longitudinal research, with currently over 50 LPS,[2] enables the effective assessment of life course trajectories and changes in societal and environmental contexts.

[1] The term 'Longitudinal Population Studies' includes a broad range of study designs: cohort and household studies involving direct contact with participants; longitudinal studies exclusively using routine records (e.g. Census Longitudinal Studies) and sequential cross-sectional studies (e.g. British National Surveys of Sexual Attitudes and Lifestyles (NATSAL)). In this study, the term primarily means cohorts and household studies with longitudinal follow-up of the same units through direct participant involvement.
[2] The Medical Research Council's 'Cohort Directory' currently lists 47 studies, although this list is incomplete.

1.6 Developing 'integrated data resources', based on maximising the use of data linkage, was seen as critical to answer diverse policy questions and to inform cross-disciplinary investigations; an observation which aligns closely with those of LPS reviews in the biomedical field (Pell *et al,* 2014, Wellcome Trust, 2017).

1.7 The LS Review also identified a substantive challenge, whereby both individual LPS, and the whole portfolio of UK LPS, were considered to potentially lack the 'representativeness' necessary to generalise findings to the wider population and sometimes to lack the breadth of data or sample size to answer questions relating to the devolved authorities or vulnerable sub-groups. The review concluded this lack of representativeness results from challenges in recruitment, changes in the population over time and from loss to follow-up in LPS where the rates of attrition vary by health, demographic and socio-economic factors.

1.8 The LS Review authors, from their international perspective, stressed the benefits of using population data to address these challenges. Specifically, they proposed the construction of an Administrative Data Spine (ADS) to support new and existing LPS across disciplinary traditions.

> "Data linkage has the potential to drive the design of longitudinal surveys, by using a suitable population register as a "spine" and linking all (or as many as possible) ESRC longitudinal data collections to it." (Davis Keen *et al*, 2017; p. 34).

1.9 The design of an ADS was not specified in the review, but to realise the intended benefits, such an infrastructure would need to comprise a population register (a list of UK residents and their contact details) and at least some attribute information. The ADS register would require maximal population coverage and the attribute data could range from a 'thin spine' of key socio-demographic variables to a 'thick spine' of comprehensive health and social records.

1.10 The LS Review also recommended that the ESRC should commission a new birth cohort in order to address the evidence gap relating to children and young adults and capturing critical evidence during early years development, because the last national birth cohort was commissioned in the year 2000. The LS Review recommended the new study should have a sampling design representative of the whole population with overweighting in devolved nations and that sampling, recruitment and subsequent follow-up could be based on a new ADS infrastructure. An 'accelerated design' could be considered where sub-sets of the cohort are sampled at older ages in order to gain insights into education and employment transition outcomes sooner (see Shlomo *et al,* 2019).

1.11 Following the review, the ESRC established the 'UK Population Lab' programme, funded under the UK Research and Innovation (UKRI) Strategic Priorities Fund (SPF). The programme includes this study, a range of additional scoping studies, a public dialogue exercise, think pieces and evidence summaries.

The Longitudinal Population Studies

1.12 LPS design decisions tend to be determined by the needs of specific research questions, available resources and disciplinary traditions: yet, all have the active involvement of participants as a common defining feature. Studies may ask volunteer

participants to agree to take part - a process bound by consent - to 'donate' biological samples and information about themselves and/or their wider family/household, and increasingly, for some of these participants to play an active role in study operations. This takes the form of an exchange, with participants providing data (with specific safeguards) in the expectation that it will be used to generate public goods through the research process. On this basis, LPS enjoy strong public support, with some studies having continued participant involvement for over 70 years and an estimated 2-3m participants across all UK LPS.

> *"These datasets have become a kind of component of our democracy… They shape how people think about people. Which is wonderful. It is part of the impact of cohort studies as most policy makers now tend to think in a life course kind of way."* (Expert contributor to this study, 2020).

1.13 To have maximum relevance to research and policy development, LPS - when considered as a portfolio of resources comprising a strategic research infrastructure - should have heterogeneous sample composition and data collection follow-up strategies that are inclusive of all population groups: including harder-to-reach groups who may be vulnerable and/or marginalised. More inclusive datasets enable researchers to gain an improved understanding of the health status, life chances and outcomes of different groups in the population and to provide evidence that will help policy makers make better decisions, target resources and deliver services of benefit to those most in need. The adequacy of current approaches to this has been challenged (e.g., Bécares *et al*, 2020).

1.14 In addition to data collected directly from participants, LPS make extensive use of 'population data', comprising routinely generated records (e.g., health, education, economic and environmental records), and increasingly, novel digital sources (e.g., consumer transactional records, social media data, phone app and internet connected sensor readings). These population data can be used to help select samples, to provide contact details for recruitment approaches and follow-up requests, to augment directly collected data, or as a source of data to help assess potential bias and to inform strategies to address missing data. The use of population data does not replace the need for data collected directly from participants; both have different scientific strengths and the two combined offer possibilities greater than the individual parts.

1.15 The UK government is investing heavily in the digital economy and considers population data science can provide more effective and efficient service provision. Organisations such as the NHS, Office for National Statistics (ONS), Health Data Research UK (HDRUK), Administrative Data Research UK (ADRUK) and the Alan Turing Institute are taking leading roles in establishing an enabling landscape for data intensive science. The resulting data and infrastructures offer new opportunities for the use of population data in LPS and consequently help frame the ethico-governance and data landscapes in which LPS ultimately sit.

1.16 Options for the centralised support for LPS are being considered by LPS funders (UK Research and Innovation, Wellcome Trust) through considerations of a Population Research UK (PRUK) initiative. The form that a PRUK may take was considered through a scoping project led by HDRUK, whose recommended model explicitly includes consideration as to how population data can be used efficiently and effectively by the LPS community.

Representativeness and Inclusivity

1.17 The LS Review emphasised the great value the ESRC LPS community place on 'representativeness' as a "primary defining feature" of a study and this is held in "high esteem" by users. This was attributed to the value of being able to accurately generalise study inferences to the wider population (typically a national or devolved/regional population) for policy making. Thus, there is a perceived need from the ESRC community for study findings to have external validity. Yet, the LS Review authors also acknowledge that maintaining representativeness is a constant challenge and that more needs to be done to ensure inclusion of people living within the devolved nations or those who are members of vulnerable or marginalised sub-populations.

1.18 Some of the characteristics used in considerations of 'representativeness' are defined through individual determination (e.g., ethnicity) rather than fact (e.g., age) and that official records may misclassify individuals. Ensuring 'representativeness' in sampling and maintaining this is therefore challenged by changing individual views and that the language and classification systems used to describe these characteristics also change. Further to this, considerations over which population sub-groups are considered to be vulnerable and marginalised will change over time and that some groups considered to be vulnerable by policy makers, service providers and academics may not identify as being vulnerable themselves either at a group or individual level.

1.19 The need for greater inclusivity relates to statistical power (e.g., whether the sample size is sufficient to consider devolved matters or outcomes in particular sub-groups) and the concern that without a fully heterogeneous sample inclusive of the harder to reach, then any study findings and consequently the policy-relevant estimates, will under-represent the extent of disadvantage amongst particular sub-groups.

1.20 In response to the LS Review, Benzeval (Benzeval *et al*, 2019) considered what was meant by the term 'representativeness', drawing a distinction between two separate but important considerations and stating that LPS should strive to:

> "*support population inferences* (bearing in mind that target populations need to be carefully defined). This means the sample members should, at least in part, be a selected from the target population with known or credibly estimated probabilities, so that the well-developed statistical methodologies for population inferences from probability samples may be brought to bear";
>
> and
>
> "*ensure sufficient heterogeneity in the sample* to enable subpopulation analyses of interest and the estimation of a wide range of associations, gradients and causal relationships between and within those subpopulations." (Benzeval *et al*, 2019, p. 6).

1.21 In contrast, Kieding and Louis (2016) describe that the importance placed on 'representativeness' is not universally held across the LPS community: with some approaches focusing on internal validity and assessing 'generalisability' through replicating findings across different studies; and others seeking to recruit a relatively homogenous population which is amenable to the burdens of long-term follow-up and to improve within-sample precision.

1.22 Three key factors reduce the contrast between these approaches. Firstly, the need to rapidly apply research findings to policy development and service planning encourages (biomedically focused) epidemiological studies to work with policy makers who will require evidence on the external validity of their findings. Secondly, the requirement for LPS to make their data widely available to help ensure value-for-money and maximum scientific value is helping develop a broader, interdisciplinary user base with an accompanying driver to meet user expectations. Finally, there is an emerging recognition that the socio-economic and health characteristics of volunteer participants may impact on their willingness for ongoing participation in an LPS and introduce bias in epidemiological studies in ways which were not anticipated (e.g., genome-wide association studies), suggesting a need for more heterogeneous samples. This issue is considered further in **Chapter 2**.

1.23 This scoping study does not consider the relative merits of different sample designs, or which design is best suited to investigate any given scientific endeavour; but it does discuss the scientific and ethical requirement for sufficient sample size in sub-groups whose data will be used in targeted ways. Rather, it will consider how this information can be accessed and used in a manner that promotes inclusive research and is efficient and effective while being proportionate and acceptable.

Trust and Social Licence

1.24 The ability to use population data within research is constrained by limitations on what methodologies are considered to be acceptable and proportionate by the public and professional stakeholders. To be seen as publicly and politically acceptable, any data use will need a 'social licence' that is achieved through setting conditions that extend beyond legal compliance and data protection and encompass ethical standards, the respect of individual rights and the delivery of public benefits. Failure to establish such a social licence has led to the collapse of proposed research data resources (e.g., the 2014 NHS 'care.data' proposal for a centralised database of English GP records) and is being seen to impact contemporary initiatives (e.g., the 2021 NHS 'General Practice Data for Planning and Research', GPDPR) (Carter *et al*, 2015).

1.25 Social licence is a critical issue for LPS given that propensity to enrol and decisions over ongoing participation will be dependent on the trust relationship between the study and its participants, which are typically framed around an understanding that donated data will be used appropriately to improve the public good. This framework for social legitimacy can be undermined by the fact that LPS operate over very-long time periods and frequently investigate new research topics or adopt methodologies that were unforeseen at the study's inception. Such changes, along with changes in wider society - including new ways of using and/or abusing data - may create challenges to maintaining participant trust.

1.26 In recognition of the need for 'social licence' within LPS, achieving inclusivity and wide public benefit is socially important. For this scoping study therefore, I am augmenting Benzeval's considerations of representativeness, which relate to the ability of LPS to draw meaningful inferences, with a third socio-ethical consideration:

> *To ensure the sample is sufficiently inclusive that it is perceived as being fair, socially just and of benefit to society; and, therefore, likely (in this regard) to maintain its social licence with the public and politicians who represent them.*

When framing the concept of representativeness in this way, it could be argued that moves to increase the use of population data *might be* viewed as ethical, legitimate and proportionate; and that not being inclusive *might* - in some circumstances - render the use of population data unethical and illegitimate. This suggests that the status quo could form a risk to the ongoing social licence for longitudinal research if LPS as a collective (rather than individual studies) were perceived as only benefiting a typically more advantaged sub-set of society.

The Impact of Changing Health and Social Environment on LPS and the Use of Population Data

1.27 This report was written during the Covid-19 pandemic, the Black Lives Matter social justice movement, and political emphasis on 'levelling up' opportunities across the UK nations and regions: all of which have highlighted health and social inequalities across population sub-groups and geographies. The Covid-19 pandemic has emphasised the need for the research community to be able to respond to rapidly changing research priorities and has revealed the value of an embedded national strategic reserve of baseline data and the competence and capacity to mobilise in an effective manner. For longitudinal research, this includes the need for regularly refreshed linked data on virus status and outcomes in order to fully interpret study data and inform policy. This suggests the need for an agile infrastructure for LPS to utilise population data in order to influence evidence-based decision making in response to emerging challenges at pace and at scale.

> *"One lesson that is very important to learn from this pandemic, and for emergencies in general, is that data flows and data systems are incredibly important. You need the information in order to be able to make the decisions. Therefore, for any emergency situation those data systems need to be in place up front to be able to give the information to make the analysis and make the decisions."* (Sir Patrick Vallance, UK Chief Scientific Adviser, 2020)

1.28 The pandemic is also shifting the views of policy makers and data owners on the acceptability of using population data in research and specifically research designed to inform policy response and service provision. Options for how population data can be used, which may have been thought unacceptable during the evidence gathering phase of this scoping study, are now possible given different levels of risk tolerance and changed perceptions around proportionality. In addition, The Health Service (Control of Patient Information) Regulations 2002, has facilitated access to hitherto hard to access datasets (e.g., English general practice records). However, we are yet to see a sustained effort to explain to the public the impact and benefits of these new ways of working with data and how they are enabling the UK to respond to the pandemic. The troubled launch of NHS Digital's GPDPR database of GP data during 2021 emphasises that the benefits of using data in Covid-19 have not changed the fundamental need for establishing a 'social licence' to ensure acceptable data science.

1.29 In response to the pandemic the UK Chief Scientific Advisor has established the National Core Studies (NCS) as a coordinated Covid-19 research programme. Within this, the Data & Connectivity NCS (an HDRUK and ONS led programme) is developing an enabling landscape of pan-UK accessible Covid-relevant datasets and infrastructure for safe and effective research. Aligned with this, the Longitudinal Health & Wellbeing NCS will use

longitudinal studies in conjunction with whole population health databases to study Covid-19 and the connections between the pandemic and associated behavioural restrictions, and health and social outcomes. To enable this, it is developing the 'UK Longitudinal Linkage Collaboration' (UK LLC) as a pan-UK resource for linking LPS to routine records and to provide a secure, trusted research environment for the analysis of these data. These data are being used by an interdisciplinary cadre of researchers using new approaches to ensuring open and transparent working.

1.30 The pandemic has driven a new professional culture, requiring work across disciplines and silos that enables effective data flows and use. Covid-19 has also increased public awareness of epidemiology, risk and the value of population data in informing decision making. For the new ways of working to become permanent, researchers will need to gain social licence and sustained legitimacy by demonstrating an ongoing need, substantial public benefits, security of the infrastructure and robustness of governance frameworks. Insights from extensive public involvement and engagement will be essential to overcome any underlying concerns about risk, personal liberty and privacy.

Scoping Study Remit and Methods

1.31 The broad remit of this scoping study is to consider the challenges identified in the LS Review relating to a potential lack of representativeness - in its broadest sense - within UK LPS and whether this means that harder-to-reach population sub-groups are disproportionately missing and are therefore not receiving the benefits research can bring. The scoping study takes an interdisciplinary view as these issues manifest across LPS and improvements to the use of population data are also likely to benefit across LPS.

1.32 This scoping study also considers how population data can help overcome the challenges which impact on selecting a study sample, recruiting selected individuals and ensuring inclusive follow-up. The evidence will help illustrate the pros and cons of any potential approach to using population data to address these challenges.

1.33 This scoping study will specifically consider four key questions (**see Panel 1**) which relate to the concept of an ADS, the functionality, proportionality and acceptability of such a resource and alternative infrastructure options or new ways of working that could deliver some or all of the suggested benefits that an ADS could bring: including innovations arising in response to the Covid-19 pandemic.

1.34 The study implementation is guided by the ESRC's Longitudinal Studies Data Strategy Core Group. It will have a UK focus and will draw on international parallels where appropriate.

1.35 Those working on the study fully acknowledge the need for, and value of, consulting the public on the development and design of new ways of working with personal information. The scoping study is feeding into the design of, and gathering evidence from, an aligned public dialogue study.

1.36 This study considers issues of what is scientifically optimal as well as what is publicly acceptable. It will focus on specific challenges inherent in LPS and how these may be resolved through data driven solutions. Any attempt to realise these should be co-developed with public/participant involvement and in consultation with stakeholders.

Panel 1: Key questions that the scoping study will seek to answer.

1. **Is there a substantive challenge to achieving inclusive longitudinal research?** The study has gathered evidence from existing UK LPS's recruitment, response and dropout rates in order to illustrate whether challenges in sampling, recruitment and retention are substantive and result in the direct or indirect exclusion of population sub-groups, particularly those considered to be vulnerable or marginalised. The evidence will help determine if any additional use of population data is justified and proportionate.

2. **What approaches are used for understanding the population coverage of UK longitudinal studies and for identifying missing populations, particularly vulnerable and marginalised groups?** The study will provide exemplar illustrations of how LPS are currently using population data to:

   - Assess sample coverage;
   - Aid strategies for maintaining inclusive samples (primarily tracing strategies);
   - To understand how population data is informing strategies to address missing data or a lack of sample representation;
   - Gather learning that could be used by existing studies and to help inform future studies and infrastructure enhancements.

3. **What are the different kinds of population data sets that are used in the UK which could inform inclusive Longitudinal Research?** The study will:

   - Identify the broad kinds of population data that exist within the UK in order to understand the possibilities of existing data resources and to help determine if a new type of infrastructure or different ways of working are needed;
   - Explicitly consider the recommendation made by the LS Review authors to construct an ADS and identify what benefits its implementation could bring.

   The study will then consider:

   - Alternative ways of realising some or all of these benefits through other methods;
   - How the different options align with other UKRI activities and the infrastructure development work being undertaken by HDRUK and ADRUK. Within this, the study will gather evidence regarding population coverage and inclusion of vulnerable and marginalised groups across the different datasets;
   - Variation across the UK nations, with a particular focus on which datasets could inform the sampling of a new birth cohort study and its ethico-legal basis.

4. **What is the ethical justification for using population data to address the challenge of inclusive longitudinal research and is this likely to have the 'social licence' needed to be acceptable?** The study will consider:

   - whether existing ethical frameworks are supportive of the use of population data for research:

   and explicitly - in order to overcome barriers to access

   - whether data may be accessed without explicit consent for this purpose. The study also considers whether LPS and the wider research community are obliged to act on this issue.

Methods

1.37 The scoping work has been informed through interviews with diverse experts and desk-based research. It has not been tasked with changing or building any resource. All evidence gathering was designed to be systematic (within the bounds of the resources available) and to consider the needs of LPS supported by UKRI and other UK funders.

1.38 The methods used were: 1) to send information gathering proformas to data owners (statistical agencies) in the UK nations; 2) to interview experts working in LPS, in government and in data infrastructures; 3) desk-based research conducting rapid literature reviews; 4) to commission expert reviews on the legal basis for the use of population data; and how to develop integrated data repositories at a local level using exemplars from Manchester and Bristol.

1.39 The methods used and results of the scoping project will be open and transparent to the public (see **Appendix 1** for a detailed description of the study methodologies).

Chapter 2: The challenges faced by LPS relating to the inclusion of harder to reach sub-groups.

2.1 The LS Review identified a number of “challenges” in the design and delivery of longitudinal research. This chapter considers the issues raised, to help identify whether new infrastructure and/or ways of working are proportionate and to help identify the shape, functionality and content of any required new resource or working method. Throughout this report, some paragraphs highlight (in bold type) conclusions, learning points and recommendations. These are summarised at the end of the chapter.

Sampling Approaches and the Challenges Encountered

2.2 Every LPS has a unique combination of scientific drivers and team expertise. Funders and peer review will help shape the most appropriate sampling and recruitment process, which must balance scientific design considerations with what is feasible within the available resources. All LPS designers share similar challenges over the time frame of any study: that the future uses of the study and areas of policy interest are unknown; and, that the recruited participants' health status and social characteristics will inevitably change over time, resulting in movement in and out of group membership. The LS Review suggested more flexible LPS designs were needed, balancing the needs of investigating specific hypotheses with the objective of building a useful data resource, so that LPS as a whole constitute a UK data infrastructure which offers ‘depth’ to compliment the ‘breadth’ of emerging pan-UK whole population electronic data sets built from routine data sources.

2.3 LPS share common structural issues when attempting to use population data to sample, recruit, retain and conduct long-term follow-up through record linkages and addressing missingness and bias. Challenges include data discovery, poor documentation of routine records (particularly of sources of bias and error in the collection and processing of these) and a lack of institutional memory or capacity. Barriers to data access can arise from a failure to enact linkages, in particular ethico-legal ‘soft’ barriers (e.g., a lack of clarity of what is possible, a lack of risk tolerance, a lack of perceived benefits and concerns regarding proportionality) and ‘hard’ barriers (e.g., where departmental legal gateways impact on the flow of data).

Approaches to Study Sampling

2.4 The LS Review stressed the requirement for policy makers that longitudinal research ensures adequate coverage of vulnerable and marginalised groups of policy interest, the requirement for the research to be accurate, and to be representative so that research findings can be generalised to the ‘full’ population. The scale of ‘full population’ is likely to differ depending on where the policy maker sits, for example in a local authority or health commissioning group, within a devolved authority, region, or in Westminster.

2.5 This report focuses on quantitative approaches to sampling and in the UK there are two main strategies to achieve adequate coverage. A number of the UK LPS place great emphasis on the value of sampling, recruiting and retaining a nationally *representative* study population. However, many LPS instead adopt *purposive sampling* with an aim to sample

and recruit based on a set of target attributes, and in some studies, with an emphasis on obtaining very large sample sizes.

2.6 Representativeness is seen as a defining characteristic of ESRC-supported LPS (Davis-Keen *et al*, 2017) as it enables policy makers to make valid population inferences and it helps provide sufficient heterogeneity in realised samples to enable assessments between and within sub-populations of interest. The LS Review clarified the key criteria of this approach. Firstly, those designing a study should define their target population ("of what population is the sample representative?") This is the population of all entities, hereafter, the 'population' that could be selected into the sample and is distinct from the total population (i.e., all residents of the UK) or the policy population (i.e., individuals of interest during the assessment or development of policy).

2.7 To define a target population the sample designers **ideally need access to population data with full coverage** (e.g., a register of all twins resident in the UK for a national twin LPS or the ADS vision of a register of all residents). Studies will need to establish (or "draw") a **sampling frame** which comprises the entities[3] which are available to be sampled and which ideally will be as close in terms of coverage and characteristics to the target population as possible. The LPS designers tend to select a set of key population attributes (e.g., age, health status, ethnicity, socio-economic indicators, and geographical areas associated with these) and then optimise the sampling by using estimates of means of these attributes. Sample stratification and clustering is often used to minimise the variance of these estimates (Lynn, 2009). **It is therefore necessary that sample designers have access to the population attributes of interest when establishing the sampling frame.**

2.8 A representative sample does not need to be either selected randomly from the population or to be a "miniature of the population" in order to support population inferences as statistical adjustments such as weightings can correct for biased samples (Benzeval *et al*, 2019). **It will therefore be crucial for the study managers to have access to sufficient population data to establish the weightings.** This flexibility enables LPS designers to stratify selection in their sampling frame to help ensure sample diversity and to achieve adequate statistical power in the regions and devolved nations. They may also choose to oversample groups (**booster samples**) with particular characteristics in order to build sufficient statistical power for inter and intra sub-group analysis or as a mitigation against predicted higher rates of attrition. A study will only have resources to recruit a set realised sample size,[4] meaning that any increase in the booster sample will decrease the size of the general population sample with a resulting loss in precision of estimates and associations. **This suggests that the booster sampling process should be as efficient as possible in order to minimise reductions of utility in the general population sample.**

2.9 Those contributing evidence to this study have suggested that weightings are not always used and can be poorly understood and implemented by LPS research users. This can potentially result in poor quality findings and subsequent inferences. Any increase in sampling complexity will inevitably result in an increased complexity of the weightings. Further investigation into weightings is outside the scope of this study, **but the evidence**

[3] For probability samples it is crucial that while each entity does not need to have an equal chance of being selected, all must have a non-zero chance of being selected and the manner of the selection should be such that the probability of being selected can be quantified, or at least estimated.

[4] Or more likely, to target a set selected pool thought sufficient to realise the desired sample size.

received suggests that additional training, clear documentation, software and user guides may be needed to support LPS users where weightings are in use; and there are likely to be long-term benefits in minimising the complexity of sampling strategies.

2.10 Many LPS however do not seek to sample or recruit representative samples and instead adopt *purposive sampling.* In these studies, there is a greater emphasis on internal validity (validity of inferences to the realised sample) than to external validity (generalising of within-study inferences to the total population) (Keiding and Louis, 2016). Within health and biomedical studies there is a much wider set of LPS designs, a subset of these are:

- biobanks primarily aiming for increasingly large sample sizes over time in order to detect and assess rare genetic ('omic) associations and follow-up participant outcomes primarily through linkage to routine records;
- LPS aiming to characterise their participants intensively through "deep phenotyping" at study assessment centres where feasibility and minimising participant burden are key requirements; twin studies which aim to assess the contribution of genetics as opposed to environments in relation to a given trait;
- LPS designed to facilitate "learning health systems" and "Population Health Management" through integrating their databanks into local health and social systems in order to provide rapid feedback into service planning and provision; and,
- LPS recruiting generations of families to identify the genetic basis of common complex diseases.

These differing objectives will lend themselves to different sampling and recruitment approaches such as seeking volunteers to proactively enrol (with no sampling frame being used); seeking to recruit extended family groups (potentially using ancestry records from civil registers as the sampling frame); selecting convenience sampling around a shared profession; or frequently, selecting all individuals conditional on a set of target attributes (e.g., a prospective city-based birth cohort using local midwifery and delivery records as a sampling frame). **With the exception of the volunteer approaches, these sampling approaches will tend to be optimised through the use of a 'register' of the target population for efficient and inclusive recruitment**.

2.11 The scoping study received guidance that given the future uses of the LPS and areas of policy interest are unknown at the point of study conception, then those designing LPS, and particularly general population LPS, **should adopt simple randomised selection methodologies emphasising heterogeneity in the general population sample (likely involving some stratification) and only boost sampling the minimum essential sub-groups**. See Sullivan *et al,* 2020 for discussion on this issue. This guidance was based on a recognition that the resulting LPS datasets will be utilised for a very wide range of possible research enquiries, and that oversampling on many distinct characteristics will significantly increase the complexity of the data and may lead to false inferences.

2.12 The guidance above raises the **potential for vulnerable sub-groups of contemporary interest to be followed through aligned but separate studies (for example, a cohort of children in the care system operating to a parallel cohort of children sampled from the general population).** These LPS could be quantitative,

qualitative or adopt mixed-method approaches utilising both the 'breadth' of large population quantitative studies and 'depth' of qualitative approaches. This will allow for efficiencies (shared back office infrastructure, fieldwork and communication management), the potential for prospective harmonisation (through sharing a core questionnaire set) and joint analysis with the general population sample where appropriate (e.g., wider population 'control' samples). **Importantly, this also allows for tailored fieldwork, engagement strategies and research programmes which are relevant and specific to the particular sub-group. These should be jointly developed with sub-group representatives and policy makers working in that area.**

2.13 While it is beyond the scope of this scoping study to explore methodologies in detail, it is important to note the role of intensive "fieldwork" based approaches to sampling used within health and research studies in general (Bonevski *et al*, 2014) and also qualitative longitudinal research (e.g., Neale, 2012). In these studies, formative research is undertaken to understand the population of interest, their lived experiences and the challenges they face (including those of policy interest); insights from this are then used to inform the compilation of a sampling frame, which can be constructed by fieldworkers operating within the community of interest and/or through the assistance of community gatekeepers and family members. Aspects of these approaches have been seen in quantitative LPS when seeking to recruit marginalised individuals (e.g., the Life and Living in Advanced Age Cohort Study in New Zealand (Hayman *et al*, 2012)) or to engage and retain populations in marginalised groups (e.g., Born In Bradford's intensive fieldwork approaches to developing relationships and data collection with their Roma population).

> *"To tackle the challenges of research with socially disadvantaged groups, and increase their representation in health and medical research, researchers and research institutions need to acknowledge extended timeframes, plan for higher resourcing costs and operate via community partnerships."* (Bonevski *et al*, 2014).

Appendix 3 provides summaries of the sampling approaches across different UK LPS case studies.

The Value of Representativeness and Heterogeneity in Realised Samples

2.14 The 'orthodoxy' of the above nationally representative approach does not extend to biomedical studies (Lynn, 2015), which have argued that a study need not necessarily be representative in order to draw valid inferences about causal relationships and that it would be preferable to sample purposively the groups of immediate relevance to the research question and those willing to undergo potentially burdensome assessments (Rothman, 2013). External validity would be assessed through replication, and by generalising findings to other LPS populations or sub-groups, rather than seeking representative samples and external benchmarking (see Benzeval *et al*, 2019 for a description of assessing representativeness).

2.15 Very large scale LPS aiming to assess genomic and disease phenotype associations, such as the National Institute of Health's Precision Medicine Initiative Cohort or UK Biobank, can be characterised as non-representative samples of potentially 'healthy volunteers'. It was anticipated by those designing these studies that these designs would enable unbiased assessments of genetic associations on the basis that these were unlikely

to be associated with selection or attrition or confounding by social position. Therefore the pragmatic decision was to focus resources on obtaining large samples sizes through purposive sampling. Recent evidence suggests however, that individuals with genetic risk for disease and related phenotypes can be less likely to participate, leading to a potential bias (Martin *et al*, 2016, Munafo *et al*, 2018, Taylor *et al*, 2018).

2.16 It is outside the scope of this scoping study to weigh the relative merits of these approaches (which has been debated extensively elsewhere[5]) and it remains probable that both will continue to be used. However, **the evidence and debate suggests prioritising sample heterogeneity - to a level reflecting heterogeneity in the target population - in a realised sample in both approaches.** Such an approach should make due consideration of the need to balance oversampling specific population groups (an issue discussed below) with risks relating to introducing (through complex sampling) considerable variations to selection probabilities which may reduce power for other analyses and risk the flexibility of the study to respond to emerging priorities. For this study, it is also important to reflect that those adopting either approach will likely be accessing and using the same data sources and for similar purposes: **suggesting centralised resources for population data based sampling and follow-up could have interdisciplinary benefits.**

Evidencing the Impact on Inclusivity

2.17 The objective for all studies is that their realised sample (those enrolling and participating) is as close as possible to the target population (the ideal population for the scientific purpose). Individuals in the target population may become excluded or under-represented at the point of studies compiling the sampling frame, when using the information in the sampling frame to invite individuals to enrol and during the enrolment process. **All UK LPS considered in this study were found to exclude or under-represent at least some sub-groups of the population when compiling their sampling frame.** This is a reflection that sampling is complicated, that the available information is not aligned with the ideal needs of the study, or that the information required is inaccessible. **Without the availability of a fully comprehensive and annotated register, sampling and recruitment contact will inevitably have at least some limitations.**

Processes Driving Exclusion/Under-representation

2.18 ***Design choice exclusions:*** **where particular groups are excluded from the studies target population or where factors mean that design choices are less effective for particular groups.** These may result from the methodological design features of any study. For example, Avon Longitudinal Study of Parents and Children (ALSPAC) only included individuals within a specified geographical catchment which inevitably had a demographic profile that was not representative of the national population. Northern Ireland Cohort of Longitudinal Aging (NICOLA) and UK Household Longitudinal Study (UKHLS) excluded individuals if they lived in residential accommodation, given that the studies sampled households. It is an interesting dimension of longitudinal research that such

[5] See comment and debate initiated by Davey Smith and Ebrahim in International Journal of Epidemiology (2013, vol 42, 1012-1028); and a reflecting set of comment and debate initiated by Goldstein in Longitudinal and Life Course Studies (2015 Volume 6 Issue 3 Pp 447 – 475); and a review by Keiding and Louis, 2016.

'exclusion' has a temporal dimension: for example, studies may develop samples of individuals in residential settings as children are taken into care or older participants move into residential care. It is also the case that due to issues of methodological feasibility, some choices designed to increase heterogeneity will be less effective for some population groups than others: for example UKHLS's area-based ethnic-minority boost sample focused on non-white minority groups as some groups, e.g., Roma, have relatively small total populations who are widely distributed (Berthoud *et al,* 2009).[6]

2.19 ***Stratification and oversampling by areal unit:*** **Where particular areas are selected based on the characteristics of the areas' population.** Typically, stratification and oversampling are used to encourage (or force) heterogeneity into the sample or to protect against anticipated low enrolment rates or high attrition rates. For example, the Millennium Cohort Study (MCS) oversampled areas with high populations of ethnic-minority residents. Stratification, oversampling and clustering approaches would ideally use individual-level data, but rather tend to use aggregated population data characterising an areal unit (e.g., a postcode sector), because individual-level information either does not exist or is inaccessible. This leaves the process exposed to issues relating to the phenomenon where the aggregate characteristics of the areal unit do not necessarily reflect the characteristics of each individual/household living in the area, e.g., not all individuals living in a disadvantaged neighbourhood are themselves disadvantaged (this phenomenon is well discussed in the literature relating to **area-based ecological fallacy**, e.g., Openshaw 1984). This can result in:

(1) a loss of study efficiency as the oversampling is not targeting solely the sub-group of interest, and while this can be addressed through fieldworker screening questions it remains inefficient in resource terms. Given budgetary constraints, any increase in sample sizes in sub-groups to compensate for imprecise sampling information is likely to mean reduced sample sizes in the general, random population: the population which may contain important, unobserved, sub-groups of future policy interest;

(2) a reduction in the heterogeneity of the realised sample as members of sub-groups living in areas of high population density of that group may have different life experiences and circumstances than those living in areas of low density. There are noteworthy examples of existing fieldwork practice controlling for this risk by also collecting sub-group specific data on eligible individuals selected through full population probability sampling: for example, UKHLS fieldworkers administered the additional survey questions allocated to the ethnic boost sample to ethnic minorities identified through the general population recruitment (Boreham *et al*, 2012);

(3) the masking of the range of diversity of circumstances amongst the sub-group of interest which could hinder the ability to tailor specific recruitment approaches and incentives or result in members of the sub-group of interest with different characteristics being under-represented in the recruited sample (e.g., where individual level data could be used to stratify within sub-group oversampling, then those with characteristics known to be associated with non-response or challenges with recruitment could be allocated a greater number of fieldworker contact attempts).

[6] It is important to note that no ethnic group – including Roma – were excluded from the UKHLS study; the point being made is that not all methodological interventions are equally effective across all population groups.

This may introduce selection bias into strata or sub-group booster samples and the impact of this may not be fully recognised as coverage and follow-up assessments are also often based on area indicators. Given that mitigating selection bias using statistical weighting relies on the selection probabilities being known - which they typically are in these cases, but only to a certain precision and breadth of factors - the adoption of richer and more precise individual-level sampling frames would enable greater accuracy in modelling effect estimates and adjustments, in addition to increased recruitment efficiencies.

2.20 ***Barriers in data access.*** Access to the optimal data can be restricted due to limitations in how data could be shared (e.g., the legal basis, or the interpretation of the legal basis) or a lack of willingness to share. This can lead to inconsistencies in decision making. For example, MCS were permitted to use information from Child Benefit records for fieldworker visits after providing an 'opt-out' mailing (Plewis *et al*, 2007), yet LifeStudy (**see Panel 2**) were only allowed to make a fieldworker approach using information from Birth Registration records following an 'opt-in' response (Clements & Gilby, 2015). **This LifeStudy opt-in pilot study demonstrates that while sampling from Birth Registers via statistical authorities is technically feasible, it emphatically demonstrates that an 'opt-in' methodology is not effective.** This 'consent for consent' issue is discussed later in the report (see 3.19).

Panel 2: LifeStudy Opt-In Pilot (adapted from Clements & Gilby, 2015).

The LifeStudy birth component pilot was designed to test a pan UK sampling and recruitment contact strategy. The UK statistical authorities (ONS, NRS, NISRA) were commissioned to sample mothers with babies aged five to seven months from Birth Registers, although NISRA were unable to do this within the pilot timeframe. There is a legal requirement for all babies born in the UK to be registered within 42 days of birth in England and 21 days in Scotland. Data were made available two months after the last date in the birth month (e.g., the record of a baby born on the 1st January would be made available at the end of March). Coupled with the registration time window, this meant a potential lag from birth to data availability of >100 days. Due to 'data protection' concerns, the statistical authorities insisted that this should be done on a 'opt-in' basis whereby the details of the selected sample could only be passed to the study/fieldwork agency with the mothers' consent. This 'opt-in' requirement resulted in a low overall response rate (two mailing varieties were tested, one had a 15.4% net response, the second 18.7% net response) and some evidence of bias (responding mothers were older than population averages and more likely to report having a partner).

2.21 ***Data Time Lag:*** there will frequently be a delay between the event of interest (e.g., booking for maternity care, the birth of a child), the recording of that event and then that record being processed and made available for reuse (see Panel 2 as an illustration of this). This can result in important groups being missed, for example parents suffering neonatal child death, or where more mobile groups move on from their sampling address prior to recruitment contact.

2.22 ***Sample Filtering:*** data owners filter sample lists to remove individuals for safeguarding reasons (e.g., victims of domestic violence, adopted children) or National Security concerns (e.g., the Prime Minister or members of the Royal Family). The

introduction of the National Opt-Out scheme in the NHS in England has resulted in >2% of the population setting an opt-out flag which is likely to bar their selection for research studies.

2.23 ***Changes in coverage in datasets:*** routine record keeping and data systems are in a constant state of flux meaning some variables are added while some cease; equally **c**hanges in government policy may impact on the nature of these registers by reducing coverage. For example, MCS used the child benefit register as a sample frame when this was a universal benefit, it is now however means tested which may reduce its utility and coverage for future sampling as it will exclude families with wages over a certain threshold.

2.24 Theoretically, but not observed directly in this study, exclusions could also be introduced through ***Record linkage issues:*** these can be introduced where there are systematic differences in the completeness, quality or composition of the personal identifiers used to link data sources together. This will likely impact particular sub-groups whose circumstances result in significant changes in their personal identifiers or whose identifiers are less likely to be recorded in database (Bohensky *et al*, 2010), such as sub-groups with no fixed or highly variable address information, following marriage breakups (where name and address may change), individuals going through gender reassignment (whose gender and name are likely to change), those converting to a new religion (who may adopt very different names).

2.25 Patterns of exclusion are likely to vary across the four UK home nations due to differences in data capture, recording and access, and in any one nation because different data sources are used.

The Case for Granular Individual-level Data in Sampling and Recruitment

2.26 The barriers described above suggest there could be meaningful improvements to sampling and recruitment arising from using individual rather than area level indicators. This would **enable a more granular selection of individuals to recruit (thus encouraging sample heterogeneity) and more efficient fieldwork targeting.** This study found some evidence of this approach in existing UK LPS sampling using opt-out contact protocols: notably the Study of Early Education & Development (SEED) cohort which used Department for Work and Pensions child benefits records to individually identify eligible sample families, assign these to different socio-economic strata and then to select areal based clusters from this frame (Speight *et al*, 2015). **The potential for cost-efficiency is dramatic, for example: 98% of SEED households sampled using individual-level data were eligible for recruitment in comparison with a 90% eligibility rate in the UKHLS general population sample (selected using area-based indicators) and only ~23% in the ethnic boost sample.**

2.27 This approach may bring additional efficiency benefits if it could be based on access to 'live' information: for example, maintaining an-up-to-date sampling screen with current contact details, mechanisms to manage in- and out-migration, being able to manage change in health status (e.g., births and deaths). A recruitment management system could also aid inclusive recruitment through **dynamic fieldwork monitoring which could swiftly identify challenges arising with recruitment within specific sub-groups or to allocate 'replacement' individual in the result of non-enrolment of an initially selected**

individual.[7] This monitoring could help channel resources in terms of fieldworkers or engagement approaches

2.28 The scoping study received indications that study sampling would be based on individual level data if it were accessible. However, barriers to access largely prevent this. The creation of the 'NHS DigiTrials' framework provides an illustrative example as to how samples can be selected from within patient record databases and through linkage and 'flagging' on the patient register can then be used for recruitment approaches and long-term tracking and tracing and outcome follow-up. The ORION-4 trial provides an illustration of this and a precedent for the transfer of patient personal identifiers to a university for postal recruitment approaches[8] in England, Scotland and Wales.

2.29 The NHS DigiTrials system is manageable as all information is selected from within a single data source: in this case, centralised English hospital records linked to the English patient register (and Scottish and Welsh equivalents operating in parallel). This is effective for trial recruitment as eligibility is determined by health status (indicated by health codes within the record). This may not be suited for selection into LPS where selection may be based on health and/or socio-economic and demographic information; the latter of which is likely not recorded in the health record. **This would be addressed using an ADS approach but could also be addressed through a minimised flow of information specific to the purpose if barriers to data sharing could be overcome.**

2.30 To address this we can look to the field of privacy-preserving data science, where barriers to the access and use of personal individual-level data are overcome through ensuring anonymity in data processing and applying robust safeguards. To date, the scoping study has not found any evidence of these being used in sampling and recruitment, yet this approach may realise new opportunities to use routine information to ensure sampling and recruitment is more precise and efficient. **To support future consideration of this, an illustrative and outline framework for 'Privacy-Preserving Sampling and Recruitment' has been developed in this study (see Chapter 7).**

Response Rates and Evidence of Attrition Bias

2.31 **All LPS suffer from declining rates of active participation over time and there is some evidence that the extent of this is increasing over time (Watson and Wooden, 2009).** Attrition can be a factor of predictable outcomes (participant out migration, participant death), loss to follow-up (through loss of contact, non-response due to participation fatigue or life circumstances) or from study withdrawal. However, comparing rates of attrition across studies is complicated by differences in study designs, sample characteristics (including the addition of boost samples), the timing/method of follow-up strategies and differences in reporting and lies outside of the scope of this study.

[7] For example, the individuals could be stratified and then randomised (likely within specific areas for fieldwork efficiency). A primary selected case could be chosen at random and matched to a number of reserve selected cases. In the event of failing to trace, make contact with or enroll the primary candidate then the fieldworker would be allocated a matched replacement from within the same strata. This later point warrants much more detailed consideration as it could introduce unobserved bias that would make subsequent inferences and statistical adjustments more complex. It would also need to be subject to a cost/benefit assessment.
[8] See ORION-4 Data Protection 'privacy notice'.

2.32 Studies use a very wide variety of techniques to address non-response and attrition; with systematic reviews on this subject identifying the importance of incentives (Booker *et al*, 2011) and reducing the burden of participation (Teauge *et al*, 2018). Many studies also reported developing 'engagement and retention' strategy documents which planned different retention approaches for different population groups (e.g., Benzeval *et al*, 2019**)** and there is an active network on this topic managed through the CLOSER consortium.

2.33 It is noteworthy that the study found evidence of two well established cohorts - the Dunedin Study (New Zealand) and eRisk (UK), which share Principal Investigators - which report consistently very high response rates over long periods of follow-up. The investigators attribute this to the value of engagement and communications to build a strong study-participant bond, that the studies offer substantive support and compensation to facilitate follow-up (e.g., offering what are effectively 'travel agent' services for participants to fly back to New Zealand) and that - to take account that non-take-up of an invitation to participate at any single time point does not rule out willingness to participate at a later date - these studies regularly ask participants if it would be convenient for them to take part in study follow-up and then slot them into the currently running 'sweep' rather than seeking a binary consent/dissent for follow-up at the first contact.

Attrition Bias

2.34 Missing data resulting from attrition results in loss of statistical power and can result in bias. Complete case estimates of exposure-outcome associations will generally be biased if the mechanism which resulted in the missingness depends on the outcome of interest. (Cornish, 2020). **The scoping study found strong, consistent evidence across studies that those lost to attrition in LPS are systematically different in terms of health status, behaviours and life circumstances than those who continue to participate.**

2.35 Typically, those who are attritted are more likely to be: male, have lower educational attainment (or a child with parents with lower levels of attainment), socio-economically disadvantaged, ethnic minorities, younger and older adults and those at greater risk of ill-health. Differential rates of attrition may result in biased estimates of inequalities, and the extent of any bias may worsen as participation rates decrease (Howe *et al*, 2013).

2.36 There is little consistent reporting of response rates in vulnerable and marginalised groups. This is not surprising given that membership of these groups can be a 'state' in which individual's move in and out of; that defining 'vulnerable' and 'marginalised' can be challenging (see **Chapter 6**); and that when linked with attrition, it is challenging to access sufficient data to define group membership.

2.37 There are indications for some particularly vulnerable sub-groups that rates of attrition are particularly pronounced. An illustrative example relates to the follow-up of looked after children - a particularly vulnerable group who are known to be systematically underrepresented in research:

- In the ALSPAC birth cohort study, Teyhan (Teyhan *et al*, 2019) linked index participants to Children Looked-After (CLA) Data Return and Children In Need (CIN) Census, identifying high levels of attrition in the families of the 346 index children identified either as 'children looked after' (in public care) or 'children in need' (social

services involvement); and finding high rates of attrition resulting in there being little data available beyond infancy for this group. **This linkage strategy illustrates the value of linkage informed analysis, but the potential for this in ALSPAC is limited by gaps in coverage in (accessible) national records.**

- In UKHLS, the study is capturing information on fostered children (and other vulnerable groups) as they 'flow in' and 'flow out' of the study over time. They observe that while the study may have only relatively small numbers of participants in these groups at any one time, cumulatively, sufficient sample size may be achieved for analysis (Borkowska 2019). However, this strategy does not enable follow-up of those leaving the study (i.e., those being taken into care) and the extent to which this is likely is not quantifiable to the study as it is an unobserved outcome; and it also does not allow full consideration of the trajectories of those in the foster system as data capture will be transitory and longer-term outcomes likely not captured.

2.38 These examples illustrate the challenge faced by studies: it is intuitive to consider that the factors associated with a child being taken into care are also strongly correlated with a family's ability, and possibly willingness, to engage in research and thus the fact of being taken into care will not be observable to the study. **Retrospective and prospective linkage to population records could help address some of these challenges and suggests the need for regular refreshes of data to identify critical events as well as close working relationships with key stakeholders (e.g., social service providers) and Patient Participant Involvement & Engagement (PPIE) to inform strategies.** It is also important to note that neither study has a specific mandate to follow-up children taken into care, nor likely the resources or capacity to intensively target follow-up across all vulnerable and marginalised sub-groups. This suggests some degree of pragmatism but also the importance of a coordinated approach to prioritisation and allocation to studies with designs most suited to the task (e.g., in this case a regional study with close links with schools, such as Born in Bradford, whose 7-year follow-up is conducted in partnership with Bradford's schools and where linkage to social services data is in place [Bird *et al*, 2019]).

Use of population data to assess and address attrition bias

2.39 Studies identify the patterns and predictors of attrition through comparing response status against participants' baseline characteristics (internal benchmarking) and through national census and aggregate records data (external benchmarking). For example, Lynn and Borkowska (2018) have used external benchmarking to demonstrate that UKHLS is broadly similar to the UK population but with some small underestimations of young adults and some minority ethnic groups. Internal benchmarking is conducted using individual level participant data, whilst external benchmarking tends to compare aggregated study data.

2.40 Studies are increasingly conducting both internal and external assessments through the use of linked participant data. The ALSPAC study, through its Wellcome Trust funded Project to Enhance ALSPAC through Record Linkage (PEARL), had an explicit objective to demonstrate through exemplar projects the different ways in which linked population data can be used in this manner, and therefore provides a useful illustration (**see Panel 3**).

Panel 3: Using linked population data to inform assessments of attrition and bias within the ALSPAC birth cohort study.

- Boyd (Boyd *et al*, 2013) linked ALSPAC index participants to the National Pupil Database and accessed (identifiable) data on the ALSPAC sample and (de-identified) reference whole-population data for English pupils in the same academic years. This was used to benchmark participant characteristics against the national population from which they were drawn, and to assess the impact of attrition. This illustrated that the enrolled sample was broadly comparable with the national sample (with differences reflecting ALSPACs regional catchment) but that those lost to attrition were more likely to be male and living in low-income households;
- The example from Teyhan (Teyhan *et al*, 2019) described above used linked education records as a source of missing outcome data and also used unlinked records to form comparison groups for similarly aged children who were ever looked-after in England and for those in the ALSPAC recruitment area. This external benchmarking allowed national and regional policy makers to draw inferences on the external generalisability of the findings;
- Cornish (Cornish *et al,* 2021) investigated whether child and adolescent outcomes measured in linked education and primary care data were associated with participation while accounting for baseline factors. They found that after adjusting for socio-economic disadvantage (already known to be predictive of attrition), attrition was still associated with outcomes including lower school attainment, lower general practitioner consultation and prescription rates, higher body mass index, special educational needs (SEN) status, not having an asthma diagnosis, depression and being a smoker;
- Mars (Mars *et al,* 2016) linked ALSPAC self-reported survey data on self-harming behaviours with linked secondary care (hospital admissions and Accident & Emergency records) and found the prevalence of self-harm leading to hospital admission was higher in questionnaire non-responders than responders (2.0 vs.1.2%) and that hospital attendance for self-harm was underreported by responders.

2.41 Researchers are also using linked population data to inform statistical approaches to address missing data. Here, linkage can be used to access proxy variables which have some correlation to the missing data (e.g., accessing the same variable from a different source; alternative sources of information with similar constructs or variables which are strongly correlated with the missing variable). For example:

- in ALSPAC, Cornish (Cornish *et al*, 2016) used linked educational records to examine the missing data mechanism for intelligence quotient (IQ) and provide useful auxiliary variables (Key stage attainment outcomes) for multiple imputation, leading to reduced bias in estimates of the association between breastfeeding and IQ;
- Gray (Gray *et al*, 2019) have demonstrated a novel imputation mechanism - using the Scottish Health Survey and Scottish secondary health records - illustrated using

an example designed to address potential non-response bias in survey derived drinking prevalence estimates. To do this they used linked survey data to estimate the differences in demographics and hospitalisation between respondents and non-respondents, then generated synthetic observations and imputed drinking prevalence to these synthetic non-respondents. These synthetic respondents were then used to adjust the analysis and produce estimates which explained the initial survey data being patterned by non-random missingness.

2.42 Record linkage is also being used for tracing purposes in order to re-establish contact with participants lost to follow-up. For example, in preparation for their 24-year follow-up, the Next Steps study traced participants using linkage to the English patient register, through the electoral roll and using other public databases and commercial tracing software (Bailey *et al*, 2017). Studies (in England and Wales) are able to seek a set-aside of Confidentiality requirements for using confidential patient data in this way through Section 251 provisions of the NHS Act 2006. **This means that tracing using health records does not necessarily need prior consent.**

2.43 There is increasing awareness that amongst the 2-3m LPS participants in the UK there are some 'serial participants' contributing to multiple studies, e.g., it has been found that 547 women in National Study of Health and Development (NSHD) are also participants in the Million Women Study; and that 937 ALSPAC mothers are also participants of UK Biobank. These overlaps have been used to assess data quality (e.g., Cairns *et al*, 2011) but this study has not found any examples of where this has been used to help understand the patterns and causes of attrition or potentially inform statistical adjustments. **This is an underexplored area which would benefit from consideration.**

2.44 The above approaches require access to linked records on attrited participants: which may introduce barriers based on legal and ethical concerns as this will typically involve a breach in confidentiality if the individual did not expect their data to be used in this way; which is an issue explored in **Chapter 3**. Relevant to this discussion, is that attempts to base linkage strategies around seeking *retrospective* consent will not fully address missingness as it will experience similar patterns of differential non-response as wider study follow-up. Studies will therefore need to utilise legal mechanisms to set aside the duty of confidentiality or to take advantage of privacy-preserving methodologies. These routes permit the use of data without explicit consent and therefore enable assessment of bias introduced by differential recruitment or attrition patterns. **There is a clear requirement for using population data in this way in both new and existing LPS.**

Key Learning & Recommendations

(1) The study identified extensive methodological good practice for integrating considerations of inclusivity into sampling, recruitment and follow-up. **A strong and universal commitment to inclusivity was found amongst those interviewed.**

(2) There is a substantive challenge regarding the inclusion of vulnerable and marginalised populations in LPS. Studies face barriers in accessing the optimal data (a fully comprehensive sampling frame) for sampling and recruitment; and vulnerable and marginalised sub-groups are lost to attrition at higher rates than other population groups.

(3) This study found no evidence describing how inclusive the portfolio of UK LPS are; meaning that some groups may be consistently under-represented or missing.

(4) Barriers to accessing individual level data have resulted in studies using area level population data for both sampling and for analysing recruitment and attrition rates; although there are emerging examples of record linkage being used to assess attrition and to inform statistical approaches to deal with missing data using individual-level data.

(5) The use of area-based indicators, particularly in sampling and recruitment, could leave studies exposed to recruitment inefficiencies (the selection of those outside the sub-group of interest) and potential loss of heterogeneity resulting from an inability to assess the breadth of individual circumstances within an area (which lie outside the observed characteristics) and target recruitment and subsequent assessments and adjustments.

(6) To implement complex over-sampling of specific sub-groups in a new study may likely impact on the flexibility of that study to investigate future areas/groups of interest, due to overcomplicating the sampling design and thus hindering accurate inferences or through reducing power in the general population/other sub-groups.

(7) The use of population data in sampling, recruitment and retention will only partly help address the challenges described above: sustained and intensive community engagement and co-development, the building of trust relations, the promotion of the benefits of longitudinal research and good fieldwork techniques have been found effective in other research designs and there are notable examples of good practice in some LPS.

(8) There are increasing examples of how linked population data can be used to assess attrition, provide tracing information and inform approaches to addressing missing data.

Recommendations

(1) **LPS should consider the use of individual level indicators in both sampling and recruitment and the evaluation of patterns in recruitment.** Barriers to accessing these data may be addressed through adopting rigorous safeguards and new ways of working: the NHS DigiTrials and the proposed model for 'Privacy Preserving Sampling & Recruitment' described in this report (**Chapter 7**) provide a potential means to do this.

(2) **LPS should seek to limit the complexity of any sampling design in any general population study in order to maximise future flexibility and ease of interpretation.** Potential options for more inclusive LPS would be stratifying selection on broad individual level indicators of vulnerability and adversity (discussed in **Chapter 6**) or that particular vulnerable and/or marginalised sub-groups of interest could instead be studied efficiently through separate, but aligned, parallel studies.

(3) **LPS funders should place sufficient emphasis on high quality fieldwork to build sustainable relationships, understanding and trust with vulnerable and marginalised groups.** This is a long-term and resource intensive endeavour which needs a sustained and consistent commitment. **This will require funding and (given an inevitably fixed budget) the requirement for this may involve trade-offs against other study aspects, will lengthen project delivery timeframes and would challenge the short-term funding award model. The proposed PRUK could include a central and sustainable mechanism for this.**

Chapter 3: The 'social licence' to allow population data for inclusive research

3.1 A consensus has emerged within the UK Population Data Science[9] community that in order for population data to be successfully utilised in research, it will need to possess a 'social licence' (Carter *et al*, 2018).

3.2 The social licence theory suggests that activities which lie beyond generally accepted norms can take place given adherence to necessary conditions. Carter argues that these conditions will extend beyond legal compliance and information security practice. Rather, in the case of data intensive research, that acceptability is based on perceptions that involvement in research is voluntary, is governed by values of reciprocity, non-exploitation and expectations that involvement will lead to public good benefits. Where this is not perceived to be the case, then the legitimacy of data science initiatives may be called into question: the failed NHS Care.data centralised primary care database programme is an example of this.

3.3 The UK longitudinal research community has long held such a 'social licence'. Millions of UK residents have actively chosen to participate in longitudinal research and many report great satisfaction in doing so. The studies are seen as a means to make a positive contribution to society. The trust participants place in LPS is arguably the research community's greatest asset. However, there is potential for the LPS social licence to be strained as studies implement strategies to make secondary use of participants' routine records and to capture data through 'novel' digital and connected mechanisms. It is vital that studies understand public/participant views and accommodate expectations relating to rights and safeguards in their research designs.

3.4 The overriding issue that sets the work of LPS apart from wider whole population data science is that LPS inherently require *identifiable* data for their operation. This is unavoidable given that LPS interact directly with their participants over long periods of time and maintain databases of participant identities to do so and that many types of data are identifiable at the point of capture. Where possible, this is explicitly described to participants and any study is conducted on a consented basis (part of the traditional grounding of the LPS social licence). However, the 'challenge' outlined in the preceding chapter suggests that **any requirement for explicit consent to effectively sample and recruit new LPS volunteers, or to establish new uses of routine records in existing studies, is likely to result in the increased exclusion of vulnerable and marginalised groups**. Hence the onus throughout this study has been consideration of how to use population data in ways to overcome this challenge, whilst retaining the ethico-legal basis for the studies to operate, to adhere to a duty to protect public and participant rights and to retain the 'social licence' for LPS.

3.5 Following the identified challenge, this chapter considers the legal and ethical basis for utilising population data within longitudinal research and evidence of the public's

[9] Defined as the "multi-disciplinary field aimed at obtaining population-level insights with public value by organizing, linking or otherwise integrating and analyzing data that pertain to individuals and their social, economic, biological and environmental characteristics and contexts" (McGrail *et al*, 2018). It is used in this report as an expansive term to capture and include all uses of population data within health, social sciences, official statistics and policy making.

understanding and acceptance. These considerations have particular regard to the use of identifiable data in the absence of explicit consent. The study findings highlight the permissive nature of the legal and ethical frameworks in this regard and that the public are accepting of this use of their data, albeit within important bounds. **The findings also suggest that in addition to meeting ethical and legal requirements, the public expect that studies will ensure effective sampling and inclusion of vulnerable and marginalised groups.**

3.6 Any new study must establish the basis for its social licence through the involvement of the public and key stakeholders to establish a 'contract' determining the bounds and the basis for the study's operations; and for existing LPS to do likewise to ensure the robustness of their approaches and to ensure they recognise and are responsive to change. It is important that the public (and participants once enrolled) have a 'reasonable expectation' as to how their data are being used by the LPS community, as social licence cannot be conferred if those impacted by the data use are unaware of this activity (Gulliver *et al*, 2018).

> "We must articulate a clear Social Contract, where citizens (as data donors) are at the heart of decision making." (Lawler *et al*, 2018).

The Legal Basis for the Use of Population Data for Inclusive Research

3.7 In order to process identifiable data for research purposes, public authorities (e.g., government departments, universities) must have a suitable administrative legal power - known as a 'vires' - which permits this; and, that they comply with data protection legislation and the case law of torts forming the Common Law Duty of Confidentiality. Processing, including flows of data through legal 'gateways' to other parties raises data protection and Human Rights compliance implications.

3.8 This study commissioned a legal review which considered the legal basis for an ADS for research purposes only**. T****he review suggests that established legal gateways in the form of the Digital Economy Act 2017 (DEA), the Statistics & Registration Service Act 2007 ('SRSA') and the National Health Service Act 2006 (NHS Act) could provide a legal basis for an ADS** although implementing this would involve seeking clarifications on legal points and exploration of operational practice regarding data owner approval mechanisms.

3.9 Not all of these Acts have UK coverage, and in some cases their powers are restricted to only some devolved nations, or have restrictions covering their use. Resulting from this: 1) whilst the SRSA has UK coverage, it only provides a legal gateway for birth registrations for sampling for England & Wales; the NHS Act only has coverage for England & Wales meaning alternative routes are needed in Scotland and NI; and, within England & Wales, the purpose supporting the flow of patient information under the NHS Act must be intended to benefit the health and social care system - which may restrict its use for considering non-health outcomes, and, where consent is not in place, this is subject to the NHS National Opt-Out mechanism.

3.10 The NHS Act defines patient information as being data that are related to physical or mental health. This definition can potentially extend to covering population data (e.g.,

occupation status) where this can be considered related to health status (e.g., where occupation is a risk factor for contracting Covid-19 and where the severity of Covid-19 outcomes are associated with occupation groups). The potential breadth of this definintion could provide a basis for using sampling criteria drawn from both health and health related sources (e.g., morbidity indicators and socio-economic status indicators).

3.11 The ONS and NHS Digital 'Public Health Data Asset' developed by the National Core Studies for Covid-19 research provides an **illustration for how in England whole-population health records (held by NHS Digital) can be linked with administrative records (held by the ONS).** Whilst developed to address a national crisis, the legal basis for this could be informative to the development of new ways of working with population data going forward.

3.12 The UK General Data Protection Regulations (GDPR) and Data Protection Act 2018 (DPA) contain specific provisions for research. The DPA makes a distinction between Personal Data (information that relates to an identified or identifiable individual) and Anonymous data (information which cannot be related to an identifiable individual). Whilst the determination of whether data are personal or not is a matter of fact, and is therefore binary, it has also been established that this determination is context specific: meaning that data may be identifiable (Personal Data) to some users, whilst at the same time are not reasonably likely to disclose identity to other users (not identifiable and therefore not Personal Data). The UK Anonymisation Network's "Anonymisation Decision-making Framework" provides a framework for making such a determination (Elliot *et al*, 2016). This provides the basis for secure research infrastructure - e.g., Trusted Research Environments, such as those operated by the SAIL Databank in Wales and the ONS Secure Research Service in England - which can be considered anonymous as a function of the situation they are in (Elliot *et al*, 2018).[10] This status of data being 'functionally anonymous' recognises that targeted data processing and contextual controls can effectively mitigate identification risk to the point where identification is not reasonably likely. Whilst the **absence of updated regulatory codes of practice on this point is impacting on LPS securing data sharing agreements on this basis (e.g., Boyd *et al*, 2019), the ICO have issued some relevant guidance and the UK Longitudinal Linkage Collaboration provides a recent proof of principle for this way of working.**

3.13 LPS (and the wider data science community) maintain GDPR and DPA compliance through processing data in accordance with Article 6.1(e) ('task in the public interest'), and for sensitive data, Article 9(j) ('scientific research or for statistical purposes') in accordance with Article 89 safeguards. **Evidence received by this study strongly suggests any ADS, or alternative solution short of an ADS, would also utilise these articles. This is significant as Article 89 has an expectation that data used under these provisions will be minimised to those strictly necessary to fulfil the purpose.[11]**

[10] For more detail, see this guidance note from the MRC Regulatory Support Centre.
[11] The principle of 'minimisation' requires that the data being used is necessary for the purpose of the research. For LPS research this offers flexibility as the 'minimum' data can accommodate disciplinary difference in approach: for example, a tightly defined epidemiological study could require a very few variables (exposure and outcome variables with covariates required to control for confounding); whilst a social science investigation could legitimately require a much broader data set with sufficient variables needed to explore the context in which any association is occurring.

3.14 Common Law Duty of Confidentiality (Duty of Confidentiality) governs the use of data provided by an individual under an expectation that it will be kept confidential; this connects with the concept of Privacy as a Human Right. It is likely that most population data of interest to LPS (and indeed LPS data themselves) will have been recorded under an expectation of confidentiality and where a reasonable expectation has been set that the data - where identifiable - are not shared beyond a specific set of users and use purposes. Private information should only be used in accordance with a foreseeable legal power, which pursues a legitimate aim and to the minimum degree as is necessary for that purpose. Aligned with this, the use of confidential data can be addressed through seeking explicit informed consent, utilising statutory powers which set aside the expectation of confidentiality (see 3.15) or through providing individuals with sufficient information that they would reasonably expect the disclosure to be made and to provide a means to object: meaning that, therefore, no breach of confidentiality can be claimed. For social licence to be maintained, any basis to address Common Law should be coupled with making best efforts to set a reasonable expectation for any new data use and for this to include clear boundaries and rules (as described previously, social licence theory suggests these will need to be tied to public benefits and mechanisms to uphold individual rights). For systematic new uses - such as sampling and recruitment - it would be beneficial to work towards a general understanding of this use of data within the population: this is explored late in this chapter.

3.15 From a broader perspective (i.e., beyond just the potential for an ADS), there are existing routes across the UK to access identifiable health data for research purposes without explicit consent and enable the setting aside of the Duty of Confidentiality:

- In England and Wales, the Confidentiality Advisory Group of the Health Research Authority ('CAG') consider applications for 's.251 support' to use Regulation 5 of the Health Service (Control of Patient Information) Regulations 2002;
- In Scotland, the Public Benefit and Privacy Panel for Health (PBPP) considers approvals to use specified health care data involving identifiable data, the creation of new data linkages, whether with or without consent;
- In Northern Ireland, applications are made to the HSC Privacy Advisory Committee for Northern Ireland ('PAC'). It is intended that these can use provisions in the Health and Social Care (Control of Data Processing) Act Northern Ireland 2016 which are similar in powers to those found in England and Wales: however, while the legislation has been enacted, at the time of this study's evidence gathering, the regulatory framework for using it was yet to be finalised, and applications to the cannot yet draw on this.

3.16 The public dialogue exercise aligned with this study (described later in this chapter) highlighted the public expectation for rigorous oversight of population data use in research and that this forms part of the safeguards needed to achieve social licence. The CAG committee provides an illustration of such an oversight board, and indeed would be involved in decisions about whether health data could be used for sampling and recruitment. Were an ADS - or any other method of using data sampled across data owners (i.e., health and non-health data) then an alternative oversight provision would be necessary to approve the use of data and apply suitable conditions and controls.

3.17 For social administrative records the DEA provides a pan-UK power for public authorities to disclose data for research, as long as the information is de-identified before

being received by a researcher. The DEA explicitly excludes health data; although recent examples such as eCHILD which brings together whole population education records with hospital admission records demonstrate that health and administrative records can be integrated through different legal gateways.

3.18 The legal review makes clear the role of Trusted Third Parties (TTPs) in facilitating legal flows of data within these legislations. **The role of the TTP is to process identifiers within separate data processing pipelines to the processing of attribute data and thus ensure organisational separation of these two classes of data. For sampling and recruitment, the TPP could also conduct the initial recruitment contact: explaining the study, providing a reasonable expectation for any subsequent researcher contact, and providing a means to object.** The TTP could be sited within the NHS (e.g., NHS Digital could choose to operate as a TTP on the basis of seeking accreditation as a DEA processor); by the ONS; or within Universities. Under this model, the TTP (where accredited as a DEA third party processor), could process data flowing under common law disclosure of patient data from Scotland and Northern Ireland, s.251 Support for English and Welsh patient data, and, for pan-UK non-health data using the DEA. **This could provide a basis to flow de-identified health and social data into new infrastructure supporting LPS.**[12]

The 'Consent for Consent' paradox

3.19 Researchers face the 'consent for consent' ethico-legal paradox when using population data to select a sample of individuals for potential inclusion in a research study; and then, using contact information from linked population registers, to approach individuals in order to invite them to take part.

3.20 The paradox impacts on selection and recruitment to Randomised Control Trials (RCTs) where cases and matched controls are selected by personal health or social status (Junghans *et al*, 2005) and also to LPS, where eligibility is defined by personal circumstance (e.g., being pregnant or being over a certain age threshold) or where in probability sampling information is used to set sampling strata or to over-sample individuals with particular characteristics. The paradox occurs when the (typically) de-identified information used to select individuals based on their characteristics is then re-identified through linkage to a register and the subsequent identifiable data is disclosed to the study recruitment team.

3.21 Even where the sole use of the contact data will be to invite an individual to take part – i.e., to seek their consent to enrol into a study – this could be deemed to breach confidentiality and privacy and raise concerns; particularly where sensitive data are used.

3.22 To date, studies have typically sought to overcome the paradox through pragmatic design choices. For example, using health service personnel to recruit participants to RCTs to avoid 'disclosure' before the act of consenting, or sampling in LPS at an aggregated population level to avoid the need for personal information or using legal mechanisms to set aside the duty of confidentiality. **However, these pragmatic choices have negative consequences, such as adding non-clinical burden to health service staff, or resulting in non-optimal sampling designs which can introduce bias or a loss in precision of estimates and associations.**

[12] A position clarified in UK Statistics Authority regulatory guidance.

3.23 Public views on the use of population data for contact purposes were tested in a parallel study to this in early 2020 (described later in this chapter). However, one can speculate that Covid-19 pandemic may have adjusted the parameters on this though as the rapid deployment of many different studies, a heightened awareness of epidemiology, and the value of research to policy development may have set a 'reasonable expectation' amongst the public that contact details could be released for research approaches where proportionate to the research aims of the study.

Ethical Frameworks for the Use of Population Data for Inclusive Research

"Groups that are underrepresented in medical research should be provided appropriate access to participation in research." (Declaration of Helsinki, 2013)[13]

3.24 Fundamental Ethics and Human Rights frameworks incorporate the principle that research should be fair and inclusive. The first and second World Medical Association (WMA) Declarations (Geneva,1948; London, 1949) set out the first code of medical ethics and established fundamental principles relating to the primacy of the health of the patient and the duty of confidentiality owed to the patient[14]. The subsequent Declaration of Helsinki (Helsinki, 1964)[15] established the ethical basis for human experimentation setting five basic principles where research should be:1) scientifically justified; 2) conducted by competent individuals; 3) based on a detailed assessment as to whether the potential benefits are proportionate to the risks, 4) given careful consideration where the experimentation may alter on the personality of the subject; and, 5) respect the primacy of the health of the patient. It also introduced the requirement for providing clear information describing the research, gaining explicit consent and testing the individual's capacity to consent.

3.25 The Belmont Review (USA, 1979),[16] which along with the WMA Declarations has been pivotal in shaping current perceptions of ethics and good practice, identified three fundamental principles: respect for persons, beneficence and justice. Within this, **the principle of justice relates to ensuring equity in the distribution of the benefits and burdens of research**. The Belmont Review principles specifically require fairness in the procedures and outcomes when selecting research subjects, whilst recognising that unjust outcomes can still arise, even where recruitment processes are fair and well conducted. The considerations and recommendations of the Declaration of Helsinki and the Belmont Review were shaped by concerns that advantaged populations and or those in a position of power, received the benefits of research while the disadvantaged bore the burdens.

3.26 The Universal Declaration of Human Rights (New York, 1948)[17] also explicitly refers to the rights of all individuals to benefit from research:

[13] World Medical Association Declaration of Helsinki. (2013). Fortaleza, Brazil.
[14] World Medical Association Declaration of Geneva. (1948). Geneva, Switzerland; World Medical Association Declaration of Geneva. (1949). London, UK.
[15] World Medical Association Declaration of Helsinki. (1964). Helsinki, Finland.
[16] Department of Health E. The Belmont Report. Ethical principles and guidelines for the protection of human subjects of research. The Journal of the American College of Dentists. 2014;81(3):4.
[17] Assembly UG. Universal declaration of human rights. UN General Assembly. 1948 Dec 10;302(2).

> *Article 2: Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. Furthermore, no distinction shall be made on the basis of the political, jurisdictional or international status of the country or territory to which a person belongs, whether it be independent, trust, non-self-governing or under any other limitation of sovereignty.*

The emphasis on universal inclusion therefore extends to:

> *Article 27(1): Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and* ***to share in scientific advancement and its benefits****.*

3.27 In relation to the use of population data for inclusive research, these early treaties are relevant only in a general sense (Elger, 2016), yet the principles expressed permeate contemporary ethical considerations. Updates to the Declaration of Helsinki[18] now accommodate research on identifiable human material and data, reflecting the growth of population data science and research biobanking. The 2013 Declaration of Helsinki has been expanded to 37 distinct principles, which should be considered in relation to each other. Many of these principles are relevant to LPS, in particular the following three principles, relating to the inclusion of vulnerable groups and individuals and; consent requirements for human identifiable data:

> *19. Some groups and individuals are particularly vulnerable and may have an increased likelihood of being wronged or of incurring additional harm.* ***All vulnerable groups and individuals should receive specifically considered protection.***
>
> *20. Medical research with a vulnerable group is only justified if the research is responsive to the health needs or priorities of this group and the research cannot be carried out in a non-vulnerable group. In addition,* ***this group should stand to benefit from the knowledge, practices or interventions that result from the research.***
>
> *32. For medical research using identifiable human material or data, such as research on material or data contained in biobanks or similar repositories, physicians must seek informed consent for its collection, storage and/or reuse.* ***There may be exceptional situations where consent would be impossible or impracticable to obtain for such research. In such situations the research may be done only after consideration and approval of a research ethics committee.***

3.28 These requirements can be interpreted as a development of the concept of fairness and a recognition of the difference in harms which may result from population data science in comparison with human experimentation. Concerns have been expressed by social scientists that the ethical principles established within the clinical domain, which reflect historical 'research' abuses and the risks relating to experimentation, have shaped the

[18] Which should be considered in conjunction with the Declaration of Taipei on Ethical Considerations Regarding Health Databases and Biobanks.

frameworks that are applied to research in the social sciences (Dingwall R *et al*, 2014). In response to this, the Academy of Social Scientists established statements of ethical principles (**see Panel 4**).[19] While presented from a social science viewpoint, these principles align with the concepts of rights and protections afforded to research participants within the biomedical frameworks. However, there is a fundamental recasting of the framing from protecting the rights of individuals (the primacy of the patient), to balancing these with the duty of individuals comprising society to contribute to the greater good:

> *"there is necessarily a balance to be struck between respecting the interests of individuals and contributing the most reliable and valid account of some issue to the public domain. Indeed, it might be argued that those participating in a democratic society have a duty to contribute to learning from which they and others may benefit, particularly if this is conducted in ways that minimize their personal risks."* Dingwall R *et al*, 2014.

Panel 4: Academy of Social Scientists: Five Ethics Principles For Social Science Research.

(1) Social science is fundamental to a democratic society and should be inclusive of different interests, values, funders, methods and perspectives.
(2) All social science should respect the privacy, autonomy, diversity, values, and dignity of individuals, groups and communities.
(3) All social science should be conducted with integrity throughout, employing the most appropriate methods for the research purpose.
(4) All social scientists should act with regard to their social responsibilities in conducting and disseminating their research.
(5) All social science should aim to maximise benefit and minimise harm.

(Dingwall R *et al*, 2014).

3.29 While there are health,[20,21] social science[22] and university[23] ethical frameworks which convert some or all of these principles into guidelines, the Global Alliance for Genomics and Health's Framework for Responsible Sharing of Genomic and Health-Related Data (Knoppers, 2014) is perhaps most suited to the context of using population data for inclusive research, given that it draws together all aspects of the discussion above.

"This Framework applies to use of data that have been consented to by donors (or their legal representatives) and/or approved for use by competent bodies or institutions in compliance with national and international laws, general ethical principles, and best practice standards that respect restrictions on downstream uses." (Knoppers, 2014).

[19] Generic Ethics Principles For Social Science Research. (2015). Academy of Social Sciences. London, UK
[20] Health Research Authority. UK policy framework for health and social care research.
[21] Medical Research Council. MRC Ethics Series, Good Research Practice: Principles and Guidelines.
[22] Economic and Social Research Council. ESRC framework for research ethics.
[23] UK Universities. The concordat to support research integrity. 2012.

The framework is explicitly grounded within Article 27 of the Universal Declaration of Human Rights and seeks to balance

"the duty of data producers and users to engage in responsible scientific inquiry … balanced by the rights of those who donate their data". (Knoppers, 2014).

3.30 These principles suggest that using population data to study vulnerable groups is only justified where this is likely to benefit the members and potential future members of these groups, and where it cannot be carried out with individuals who are not vulnerable. Where this is the case, then explicit consent is not necessary. Wherever practicable, the use of identifiable data should be minimised and the confidentiality of sensitive data protected through anonymisation principles (at least until a point where the ethics of using these data can be based on explicit consent). For this to have legitimacy, the research programme should have robust safeguards and the researchers should strive to communicate and explain the rationale for the use of population data to ensure the vulnerable and marginalised share in both the benefits and burdens research brings.

"We must ensure a reputation for reliability, honesty, and competency in how and why we use data. Transparency at every step is vital, if we are to maintain the social license for data-driven research." (Lawler *et al*, 2018).

Equality in Longitudinal Research

"Public bodies should place considerations of equality, where they arise, at the centre of formulation of policy, side by side with all other pressing circumstances of whatever magnitude." [24]

3.31 The Equality Act 2010 requires public authorities in England, Scotland and Wales to not discriminate against, harass or victimise any person or group on the basis of 'protected characteristics' and make reasonable adjustments for individuals with protected characteristics.[25]

3.32 Universities and some research funders (including UKRI, but excluding charities) are defined as public authorities where part of their function is to conduct research and to commission and evaluate research needs. **Therefore, LPS and LPS funders have both general and specific duties under the Equality Act 2010, including the Public Sector Equality Duty (PSED)** defined in Section 149. This requires that a public authority must (a) eliminate conduct prohibited under the Act and both (b) advance equality of opportunity, and (c) foster good relations between persons who share a relevant protected characteristic and persons who do not share it. The following sections consider the research implications of these for the LPS community in turn.[26]

[24] Stuart Bracking and others v Secretary of State for Work and Pensions [2013] EWCA Civ 1345, McCombe LJ at para 60.
[25] Protective characteristics include: age, gender reassignment, being married or in a civil partnership, being pregnant or on maternity leave, disability, race including colour, nationality, ethnic or national origin, religion or belief, sex, sexual orientation.
[26] This report does not discuss the non-research duties that LPS, Universities and Funders have.

Discrimination and Reasonable Adjustments

3.33 To comply with the Equality Act, LPS must make reasonable adjustments to ensure neither direct nor indirect discrimination against individuals takes place, for example, by ensuring study assessment centres are accessible to any disabled person and by providing appropriate support such as language interpreters and study materials in large print where necessary.

3.34 LPS must therefore consider the complexities of the interactions between study methodologies and the protected characteristics. For example, by moving to solely online recruitment or data collection in a general population, an LPS could be determined to be discriminatory on the basis that older people (with 'age' being a protected characteristic) and those with disabilities (with 'disability' being a protected characteristic) are less likely to use the internet than other population groups.[27]

Advancing Equality of Opportunity

3.35 Compliance with Section 149(3) of the PSED requires the LPS community to 'advance equality of opportunity' by having regard for the need to:

> ***s149(3)(a)*** *remove or minimise disadvantages suffered by persons who share a relevant protected characteristic that are connected to that characteristic;*
>
> ***s149(3)(b)*** *take steps to meet the needs of persons who share a relevant protected characteristic that are different from the needs of persons who do not share it;*
>
> ***s149(3)(c)*** *encourage persons who share a relevant protected characteristic to participate in public life or in any other activity in which participation by such persons is disproportionately low.*

3.36 Section 149(3)(a) could be interpreted to mean that LPS and LPS funders should have **due regard for the disadvantages linked to protected characteristics when commissioning, designing and conducting research studies**. This would include considering which groups ('persons') would benefit from the research, how they would then be included in the research and how the research findings could be disseminated and explained appropriately in order to ameliorate that disadvantage. This interpretation assists with the extended consideration of the term 'representativeness' (see 1.26) which here could refer to ***LPSs being representative of the population (from a socio-ethical viewpoint) when they enable research to inform society about disadvantages linked to protected characteristics.*** It is however important to recognise that not all harder to reach sub-groups will neatly fit into definitions of Protected Characteristics.

3.37 The requirements of Section 149(3)(b) indicate that the sampling design needs to be inclusive of any group with the protected characteristic(s) and that the methodology should adopt evidence-based best-practice in order to provide 'reasonable adjustments' to meet the groups needs. For marginalised groups with protected characteristics this could mean substantially different fieldwork approaches to those used with the wider, general population (whose protected characteristics are not suggestive of needing specific approaches).

[27] Almost all UK adults aged 16 to 44 years were recent internet users (99%) in 2019, compared with 47% of adults aged 75 years and over and 78% of disabled adults. See: Internet users, UK: 2019. 2019. ONS, London, UK.

3.38 This interpretation is explicitly reinforced by Section 149(3)(c) which suggests a requirement for LPS to 'encourage' groups sharing a protected characteristic to participate in Longitudinal Research (a component of public life), particularly where rates of participation have been seen to be disproportionately low. This suggests that 'due regard' should extend beyond rigorous methodology to 'active encouragement'. When designing research programs, this could involve a range of considerations likely to result in benefits for the specific group, differential 'compensatory' benefits for participating, different communication approaches and the need to include members of the group or aligned stakeholders in the design and oversight of the study.

3.39 In s149(5) the legislation sets out the need to 'foster good relations between groups sharing a protected characteristic and those who do not share it, which requires due regard to the need to tackle prejudice and promote understanding. **This requirement could be interpreted as emphasising the need for LPS communications (participant mailings, press-releases, social media usage etc.) and findings to accurately and sensitively contribute to wider efforts to build civil relations between groups.**

3.40 There is little guidance to suggest how the Equality Act applies to population research. The Equality and Human Rights Commission's technical guidance[28] confirms that 'research and audit' is a function within the scope of the Act, but considers this scenario "may be at least one stage removed from members of the public". Nonetheless there is a clearer indication of a duty to ensure "elimination of discrimination or the advancement of equality of opportunity" in the evidence presented to policy makers. However, while this guidance may be appropriate for research using anonymous whole-population databases, this overlooks the long-term direct relationship between LPS and their participants, which may be argued to draws back into scope the direct duties described above.

3.41 Advancing equality of opportunity does not mean that all groups within a population should be treated equally in any given research study (S149(6)). An illustration of this would be the Southall and Brent Revisited (SABRE) longitudinal study which aims to understand the different levels of health outcomes in Black and Asian first-generation migrants to the UK (Tillin *et al*, 2012) or the Million Women Study which aims to investigate whether the risk of breast cancer is associated with use of different types of menopausal hormone therapy (Green *et al*, 2019). In these examples, it is appropriate that the study would disproportionately sample, recruit and follow groups with specific protected characteristics. **However at a holistic level, if there was little or insufficient research being conducted about a group with a protected characteristic who were suffering disadvantage, then the decision not to commission such research, or to exclude or not put in sufficient measures to adequately represent such a group from a new research study, would need careful consideration and could be open to legal challenge.**

Equality Impact Assessments

3.42 The Equality Act requires public authorities to consider the impact on equality of proposed changes to policies, procedures and practices. It does not require that these

[28] Equality and Human Rights Commission. Equality Act 2010: Technical Guidance on the Public Sector Equality Duty England. London: EHRC. 2013.

considerations are documented, although documentation has been seen to be valuable if decision making is challenged (Pyper, 2018).

3.43 Equality Impact Assessments (EIAs) provide a framework to assess the impact of actions on equality and to document the 'due regard' process and the decisions that were made. This study has not found any examples of these being used within the UKRI supported LPS community, although examples have been found within UKRI[29] itself, within government commissioned longitudinal studies (e.g., the Department for Education's SEED cohort, Speight *et al*, 2015) and for the development of population data infrastructure and policies within the NHS[30] and the ONS.[31] These provide indications of how PSED requirements are considered in practice. This would suggest the designers of a new LPS would need to consider four core questions to ensure compliance with the Equality Act 2010:

1) Is the proposed scientific rationale and business case for the LPS in accordance with the Act (taking into account the business case for other relevant UK LPS)?

2) Is the proposed sampling and recruitment methodology, and the information used to inform this, in accordance with the Act?

3) Is the proposed data to be collected and the collection strategy in accordance with the Act?

4) Is the operation of the LPS in accordance with the Act?

3.44 In addition to this, the LPS funders would need to consider whether the proposed LPS design, when considered in the wider landscape of all LPS and broader population data science studies and the identified research requirements, is in accordance with their strategy for meeting the requirements of the Equality Act 2010.

Understanding of the Equality Act Requirements in the LPS Community

3.45 It is not clear that the requirements of the Equality Act and PSED are widely understood and fully implemented within the LPS community or, indeed, within the wider UK data science community. This study cannot find any evidence that the ethical review bodies tasked with reviewing LPS methods and activities include explicit reference to Equality Act requirements, or to the need to protect against discrimination, or to the advancement of equal opportunities through research. Although it is important to recognise that many bodies will be doing this routinely in their work (for example, requiring that studies implement paper-based versions of online data collection materials).

3.46 There are numerous references in the reviews examined by this study to the rights of individuals and evidence of practice that aligns closely with the Equality considerations described above. Therefore, this study is not suggesting that due considerations are not being made, rather that they are not being sufficiently recorded to provide a fully effective audit trail to demonstrate compliance with statutory requirements. This applies to faculty

[29] Equality Impact Assessment for Research England Quality-related Research (QR) and formula-based research capital funding: plans and sector engagement. June 2019.
[30] National data opt-out: equality impact assessment.
[31] Equality Impact Assessment for the 2021 Census. UK Statistics Authority December 2018.

ethics guidelines of institutions managing LPS,[32] to University UK guidance (which has been endorsed by UK government and LPS funders),[33] to the Health Research Authorities' National Research Ethics Service framework and to guidelines issued by the LPS funders[34] and UK government.[35] There are explicit requirements in all of these guidelines to observe legal requirements. However, in practice there needs to be increased awareness that this must include the Equality Act[36] and any associated codes of practice and guidance. This will then align with the University UK concordat which states the need to "clearly identify and indicate any specific codes of practice and other policies that researchers and employers of researchers are expected to comply with, beyond those that might be generally expected".

Public Views and Understanding

3.47 The ESRC's 'Population laboratory' programme commissioned a literature review of public attitudes to the use of population data in research and data science (Kispeter, 2019). This identified a range of key conditions impacting on the trust and trustworthiness upon which public support depends (**see Panel 5** which is adapted from Elias, 2021). The review found evidence that studies using deliberative methods and which provided participants information about research practices generated increased support and acceptance; or at a minimum, less concern about the use of data for research purposes.

3.48 The ESRC commissioned a public dialogue seeking insights into the acceptability of using population data in LPS sampling and recruitment and ongoing follow-up (see Elias, 2021 and Coulter *et al,* 2020). The dialogue consisted of a stakeholder workshop and **two waves of workshops comprising 100 public participants (the same participants attended both waves), in five locations across the UK** (Birmingham, Cardiff, Edinburgh, London and Newcastle). The workshops were held in January and February 2020. The dialogues specifically tested issues relating to the use of population data for *inclusive LPS research* and the challenges relating to this. Public contributors recognised the importance of **representative sampling and the challenges of addressing representation and retention issues** and gave conditional agreement **that using administrative data was a potentially efficient and cost-effective way to address the inclusivity issues**. Findings emerging from the public dialogue were:

- Effective communication between the research community and the public is paramount to the promotion of the beneficial nature of longitudinal survey research

[32] The following institution's (amongst those which host LPS and where guidance was externally accessible) faculty ethics frameworks were identified and searched: University of Bristol; University College London; University of Essex; University of Oxford; University of Southampton.
[33] Universities UK. 2012. The concordat to support research integrity.
[34] Research Councils UK. RCUK Policy and Guidelines on Governance of Good Research Conduct. Updated April 2017.
[35] Rigour Respect Responsibility: Universal ethical code for scientists. Government Office for Science, London, UK. 2007.
[36] An indication of this is provided by the University of Bristol faculty ethics framework which, in their 'research involving humans' ethical checklist requires consideration of Health and Safety, Data Protection and Prevent Duty legislative requirements but does not mention the Equality Act; however, it explicitly includes a requirement for consideration of "Special issues relating to children and vulnerable adults". It should be noted that the University of Bristol is being used to illustrate this point solely on the basis that it was the only faculty ethics guidance to include a (accessible) checklist to guide committee deliberations.

and critical to this study that this needs to **target the breadth of likely recipients (i.e., those vulnerable and marginalised);**

- That concerns around involvement in LPS are influenced by perceived security risk (e.g., theft and hacking, social stigma, loss of autonomy) and that there is **little awareness of research governance and information security practice**;
- The **need for partnership working** between the research community and potential/actual participants and community/third sector groups representing the views of specific demographic groups and/or to facilitate engagement with them;
- **The oversight of LPS security and confidentiality measures by University ethics committees is not regarded as sufficiently independent to foster public trust.**

Panel 5: Key conditions of public support for population data science.

- Research must have a clearly articulated purpose which is communicated to the specific groups whose data are being used and who may benefit in clear terms that illustrate how the purposed data use and research relates to their lived experience;
- Trust can be built through the clear communication of privacy and confidentiality issues and impacts and the extent of safeguards such as de-identification. Data security, in particular safeguarding data against misuse, and effective data governance was identified as an important condition for public support;
- Other factors impacting on trust include the respect for autonomy, consent mechanisms, where the user is based: particularly distinguishing differences in acceptability between the NHS, Universities, Statistical Agencies and the private sector, and whether the research would involve profit making or not;
- There are complex interdependencies between trust, transparency and authenticity where transparency is necessary to build trust, but trust is required in order for the transparency to be recognised as adequate;
- Considerations about autonomy and consent were linked to the public's trust in the ability of research organisations to keep their data safe and the sensitivity of the data (where data about mental and sexual health, sexuality and religion were seen as particularly sensitive);
- There is debate over the terms 'public benefit' and definitions of whom the 'public' are; with some suggestions that the scope of 'public' should be as inclusive as possible;
- There was recognition from the public that increasing scientific knowledge of benefit and that the public could benefit from greater engagement with the scientific community.
- There was consensus in the literature that the public want more 'two-way' communication about data-based research, particularly on socio-ethical implications and safeguards.

3.49 Those leading the dialogue work **recommended a strategic approach for LPS engagement with the public and participants.** Within this, studies should develop (or review) overarching engagement plans to ensure they are sufficient to build trust, communicate benefits, promote both specific and wider societal benefits arising from LPS research, to improve communication about safeguarding mechanisms (detailed recommendations are made on this specific point) and to form and publish oversight mechanisms. **They also recommended that the funders provide dedicated and sustained funding to support these engagement activities.**

Raising Public Awareness

3.50 To fulfil the needs of inclusive research, it suggests the need that such a 'strategic approach' would need to make best efforts to reach the whole population (in order to meet PSED requirements). It should make the case for longitudinal research and concepts such as 'Trusted Research Environments', 'Ethical Review' and the rationale for why opt-out approaches can improve fairness and equity in research: i.e., to raise awareness of the component blocks that can help generate social licence.

3.51 The Covid-19 pandemic provides an opportunity for this given the high-profile role of data, epidemiology and other branches of science in tackling the pandemic. However, the recent troubled introduction of the NHS Digital GPDPR database suggests there remains public and professional concern over the use of data where safeguards are not clearly communicated even within the context of Covid-19 research. Whilst the language and case studies developed by Understanding Patient Data and the ethos of campaigns such as #DATASAVESLIVES were useful in countering criticism of this initiative, these did not have the reach to counter public anxiety and misinformation and misconceptions generated in the absence of clear descriptions of safeguards and boundaries on data use.

3.52 A (longer-term) route to raising awareness across the population would be to incorporate content describing longitudinal research and more broadly, population data science, within the National Curriculum (and equivalent frameworks for specifying education provision in devolved nations) and/or other school resources. This need not require statutory change, but could be achieved by adding specificity to the *existing standards* which provide opportunities to integrate such content (Pittard, V. 2018) or through alternative channels for promoting topical issues to students (**see Panel 6**). For example, materials could contribute to mathematics (statistical analysis and drawing inference), geography (epidemiology), computing (infrastructure and safeguards as well as 'Personal, Social and Health Education' (data protection rights) and 'Citizenship' (civic duties). Pertinent examples include:

- The anti-fraud education lesson plans for KS3&4 (ages 11-16) developed by the CIFAS fraud prevention service for use in PSHE teaching and which raise awareness and explain complex issues such as identity theft and financial scams.[37]
- The data protection lesson plans for KS3&4 developed by the Information Commissioner's Office 'Schools Project' which explain data protection concepts and rights.[38]

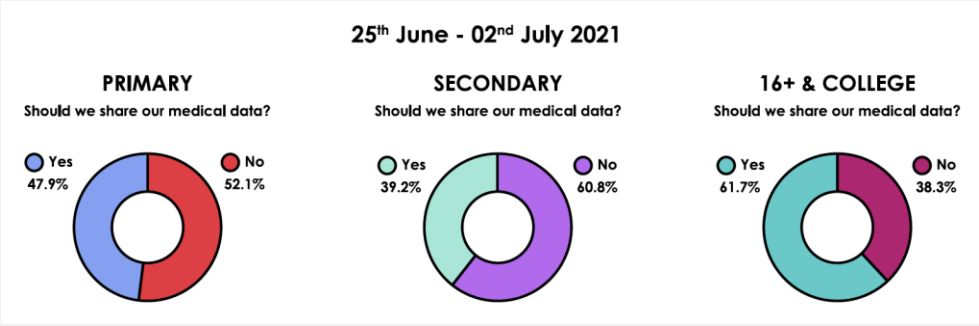
[37] Cifas. 'Anti-fraud lesson plans for KS3&4'. Distributed by the PSHE Association.
[38] The Information Commissioner's Office. 'Your rights to your information'.

- The learning materials and demonstration datasets produced by CLOSER which aim to develop capacity at an undergraduate level.[39]

The ESRC should consider options for working with education professionals, relevant stakeholders (including HDRUK, ADRUK, the Department for Education and Department for Health and Social Care) and the longitudinal and wider data science community to developing materials to raise awareness of longitudinal research and key data science concepts for delivery within secondary education and then implement this to help build the social licence for using population data for inclusive research. This could draw on existing work (such as that of CLOSER and Understanding Patient Data) and should have public/participant involvement and seek guidance from the privacy lobby and the Information Commissioner's Office in order to achieve balance and credibility. Support for promoting and encouraging schools to adopt the use of these materials would be needed in order to ensure the effectiveness of this recommendation.

Panel 6: Example school-based debating exercise in response to the proposed NHS GPDPR dataset roll out.

'Votes for Schools' (Demtech Ltd.) provides paid for teacher resources on a weekly basis. These materials are designed to enable students to consider and debate topical issues. In July 2021, in response to the proposed roll out of the NHS GPDPR, the exercise[40] considered the pros/cons of sharing health records for research.



Source: VoteForSchools.

Key Learning & Recommendations

(1) **LPS have an obligation to conduct high-quality research that is inclusive of vulnerable groups and groups suffering disadvantage(s) related to their protected characteristics and population data will help achieve these aims (a point justified elsewhere in the report).** To achieve this objective will mean ensuring there is an appropriate legal and ethical basis for the research, that the public are empowered to help shape the project and that there will be sufficient social and technological controls deployed to mitigate risks associated with participation. This will help ensure that the use of population data to inform and support LPS strategies has social licence.

[39] CLOSER longitudinal research consortium. 'Learning Hub'.
[40] A report on the exercise is available on the VoteForSchools public results webpage.

(2) Not every single study needs to be fully inclusive. Rather, **the requirement to be inclusive and equitable lies across the whole LPS community, where longitudinal research should be inclusive at a strategic level.**

(3) **The established international frameworks for research ethics recognise that population data science is generally of low risk and they make provisions for the use of identifiable data without explicit consent:** where this is proportionate to the potential benefits that can be realised and that adequate safeguards are in place.

(4) **The legitimacy of including vulnerable and marginalised groups in LPS through the use of population data is dependent on the research programme being likely to deliver meaningful benefits to these groups.** Engaging with, seeking the input from, and building trust relationships with such groups will involve a long-term commitment which will need adequate resourcing. There is public support for the use of population data in this context, but it is contingent on adequate security standards and oversight arrangements.

(5) **There is a legal basis for utilising population data to support longitudinal research, although it is complex and varies across the UK nations.** As data use by LPS is ultimately conducted on an identifiable basis (given the LPS holds the study administrative database), LPS are, to some extent, facing additional challenges compared to whole population data science initiatives using controlled and effectively anonymous information. This introduces barriers such as the consent for consent paradox. LPS will need to utilise the available tools and techniques to mitigate concerns relating to this and centralised functionally anonymous infrastructure may provide different opportunities (**see Chapter 7**).

(6) Equality Act requirements apply to LPS studies and the LPS community, across the LPS lifecycle and on an ongoing basis. At a more strategic, national level, the requirements also apply to LPS funders. LPS funders should therefore monitor LPS compliance with equality legislation, for example, through tailored metrics or by reviewing participant engagement and inclusion plans. **Those considering LPS at a strategic level will need to assess population coverage across the breadth of all LPS studies in order to identify gaps in coverage and opportunities where good coverage can address gaps in evidence.**

Recommendations

(1) **LPS must develop and implement 'Inclusion plans' and that these should be developed and refined with input from participants and members of the impacted sub-groups.**

(2) LPS funders have a legal duty to consider equality of LPS investments and should provide adequate resources and tailored performance metrics to enable individual studies to do this. **A new mechanism should be developed to consider population coverage at a LPS community level: and an outline approach for this is described in Chapter 7.**

(3) It is recommended that the ESRC consider options for **including longitudinal research and key data science concepts as a topic in schooling** in order to improve understanding of LPS and **to help set a widespread 'reasonable understanding' for the use of population data.** Non-statutory approaches, such as the development of teacher resources, provide a feasible means to help achieve this aim.

Chapter 4: An 'Administrative Data Spine' for Population Research

4.1 The LS Review authors recommended the development of an ADS to facilitate longitudinal study sampling and to enable the follow-up of participants through linkage to routine population data. This chapter explores this concept in more depth, considering what an ADS would look like, how it could work in practice and the benefits it could bring. It then considers whether such a resource would be proportionate to the challenge, and whether it would be acceptable within the UK.

What is an Administrative Data Spine?

4.2 To deliver its stated purposes, a *theoretical* ADS would have two defining attributes: firstly, an up-to-date and continually refreshed population register containing names and contact details (the population 'spine'), and secondly, attribute information about each individual represented on this spine. The requirement that longitudinal studies (or at least, the national general population studies) recruit representative and inclusive samples, suggests the spine would have maximum population coverage of all UK residents, with mechanisms to ensure inclusion of the vulnerable and marginalised. The attribute information would, as a minimum, need to include sufficient indicators to inform the sampling process.[41]

4.3 To fulfil the requirement for follow-up through record linkage, the ADS should contain the necessary attribute data or host linkage 'keys' enabling efficient linkage into diverse departmental databases. In theory, this suggests that a 'thin' ADS would be possible, containing minimum necessary data (name, contact details, the essential attributes for sampling and linkage keys). However, the likely cost of creating and maintaining an ADS would only be proportionate were the ADS to have value across disciplines and potentially other forms of population research (e.g., clinical or social Randomised Control Trials). To accommodate the breadth of these use cases it suggests that in practice a 'minimum' dataset would need to be more comprehensive, resulting in a 'thick' ADS containing additional attribute data. This could have a wide range of research uses **(see Panel 6).**

Flexible and responsive research infrastructure for emergencies

4.4 The Covid-19 pandemic provides an example of a rapidly emerging situation that demands a swift and flexible research response. LPS are well-placed to contribute to such emergencies, given that longitudinal data banks offer detailed and diverse baseline information on individuals which can offer insights into emerging complex phenomena and behaviours which are not routinely recorded within government data. Furthermore, LPS can rapidly collect new data tailored to the emergency: for example, the development and implementation of a Covid-19 questionnaire across the LPS community.

[41] The perceived 'minimum' attribute content is likely to vary by user and use case. These summary indicators could include: age, gender, ethnicity, socio-economic status, years in education, health care utilisation.

Panel 6: Theoretical benefits of an Administrative Data Spine to Support Longitudinal Research

1. a register could act as a sampling frame if it contained sufficient information to select people (e.g., a representative population sample, vulnerable individuals, a case/control sample) and to invite them to enrol;

2. individuals within a register could be clustered into households and possibly family groupings enabling different sampling (e.g., household panel studies) and recruitment opportunities (e.g., identifying new generations of existing LPS participants);

3. the register could help *trace and recontact* participants in existing studies - who were lost to follow-up and thus reduce attrition and loss of vulnerable participants;

4. the detail in the register could allow linked LPS to *assess the nature of representation and attrition bias* in their studies and to inform strategies to address this (e.g., through generating weightings, imputation models, or direct adjustment of estimates and the creation of statistical bounds on effects);

5. the register could *host persistent record linkage keys* enabling efficient and consistent linkage between people and their routine records and also between people and the natural and built environment around them;

6. given that 2-3 million people in the UK participate in LPS, the register could allow studies to quantify participant overlap across different LPS and to negate the potential harms (e.g., error introduced by non-independent samples), and benefits (e.g., data quality assessments the re-use of phenotypic and 'omic assessments) resulting from this;

7. the register could aid the public transparency and acceptability of LPS by providing an infrastructure to describe how routine records are being used in active or passive LPS and to also to host and make available individual research consent, assent, dissent status (e.g., the NHS National Data Opt-Out status);

8. the register data can be used in the s*election of matched controls* for nested LPS case/control studies. The register could provide unknown or missing matching criteria data;

9. the register data and the secure environment used to host the register could provide a 'sandbox' for government departments wanting to conduct methodological research (e.g., developing improved linkage techniques) which is not permitted within departmental data usage agreements or legal gateways (i.e., basic research rather than service evaluation or audit); yet are permitted within research gateways and agreements.

4.5 While the Covid-19 pandemic is an extreme event (in terms of its global health and socio-economic impact), it is predictable that future epidemics will occur and there have been other recent shocks, such as the 2008 economic crash, which would also have benefited from rapid research insights to inform policy intervention. This way of working would also fit with national-level situations which demand rapid insights (for example, the

ongoing high incidence of knife-crime in young adults) in addition to ongoing policy influence on issues such as health and social inequalities.

4.6 An ADS type infrastructure would assist efficient response to such social issues by providing timely flows of new data (e.g., Covid-19 test results in order to determine 'case' status), to help identify sub-groups at higher risk or in need of tailored intervention and support (e.g., involving crime or substance abuse), and to identify where populations of interest are LPS participants (e.g., to inform recall studies). It is plausible that an infrastructure which uses population data in such a way as to deliver meaningful benefits to the population or sub-groups at times of extreme need would be seen as acceptable and proportionate; although rigorous safeguards would be needed for long-term public support.

International Examples of the Use of ADS Approaches by LPS

4.7 Recently, Chambers (2020) – an LS Review author – has discussed international examples which are being developed for service planning purposes as national statistical agencies seek to move from enumerated censuses to dynamic 'census like' Population Registers. International examples include the New Zealand 'Integrated Data Infrastructure' (see **Panel 7**) and the Manitoba 'Population Research Data Repository' (Katz *et al*, 2019). However, **neither of these takes the form of the integrated ADS model described above and rather take great efforts to keep identifiers and attribute data entirely separate.** In this sense, these examples are distinct from the registry-based forms of government administration found in Scandinavia. In the absence of a central population register and the absence of universal citizen ID numbers, these population registers are compiled from information in different departmental datasets using probabilistic linkage algorithms. These are capable of delivering high rates of population coverage through leveraging the respective coverage in different departmental (and other) datasets. However, all are likely to under-count some sub-groups (e.g., the very wealthy, the most marginalised and the most mobile populations).

"the construction of a spatiotemporal 'data spine' that, through linkage to more specialized data sources, can serve as the foundation for provision of population-representative cross-sectional and longitudinal data at regular intervals, covering health, genetic, demographic, education, social welfare, and socioeconomic dynamics. It also provides the capacity to identify and to follow up key population cohorts, as well as to allow rapid increases in sample size as necessary to produce representative and highly granular 'small area' cross-sectional and longitudinal analyses." (Chambers, 2020).

4.8 The infrastructure developed in New Zealand and Manitoba has been used to deliver all the benefits suggested within the LS Review. For example:

- The Manitoba repository has been used to sample and recruit the Study of Asthma Genes and the Environment (SAGE) birth cohort and utilises a sampling strategy which over-samples children from rural, low income areas and First Nation communities (Kozyrskyj *et al*, 2009); in another example, a study of outcomes of children in care with a diagnosis of Fetal Alcohol Spectrum Disorder (FASD) eCohort was sampled from health and social records (Fuchs *et al*, 2009);

Panel 7: The New Zealand 'Integrated Data Infrastructure' resource

Statistics New Zealand (Stats NZ) have established a 'integrated Data Infrastructure' (IDI) research resource comprising routine records from government agencies, Stats NZ surveys and non-government agencies (Milne *et al*, 2019). The IDI is intended to improve service delivery and is underpinned by legislation permitting the sharing of data for research and operates within the framework of data protection safeguards based on the '5 Safes' approach.[42] The IDI comprises a population 'spine' listing those 'ever resident' in NZ and, secondly, a de-identified database of primary and secondary health care records and social provision (including: demographics; tax, income and benefits; education; justice; and civic registrations).[43] Each resident is allocated a unique encrypted identifier which can be mapped to households, providers and geographical areas. The IDI is refreshed quarterly, with the 'spine' being rebuilt from routine records on each occasion by linking birth records, tax records and visa records. However, due to data issues in the source records, and potential linkage error, the IDI has an undercount of Pacific, Asian and older Māori populations.[44]

- The IDI, Mantioba repository and the Scandinavian registries have all been used to facilitate record linkage follow-up: in the NZ Dunedin cohort (Caspi *et al*, 2016); in the Manitoban SAGE and CHILD cohorts (Kozyrskyj *et al*, 2009, Azad *et al*, 2016); and the Norwegian Mother and Baby (MOBA) cohort (Magnus *et al*, 2016) and Danish National Birth Cohort Study (DNBC) (Olsen *et al*, 2009);
- The DNBC also provides an example of using the Population Register to assess representativeness and the impact of attrition (Bliddal *et al*, 2018);
- The Manitoban FASD eCohort, the Dunedin study utilising IDI and the DNBC all leverage the possibilities found in integrated repositories drawing data from across different departments and sources.

4.9 An alternative approach is found in the ODISSEI programme in the Netherlands, which is an academic led hub developing an enabling infrastructure for social-science research in collaboration with Statistics Netherlands. In this example, Statistics Netherlands maintain the population, property and business registers and use persistent identifiers to be able to link to consenting participants health registers and social datasets (e.g., the Doetinchem Cohort Study, Verschuren *et al*, 2008); whereas ODISSEI provide the secure computing facilities, manage data discovery and access, and facilitate the linkage of survey samples to Statistics Netherlands records but do not hold any data themselves.

[42] Moffit T, McDowell A. New Zealand's Integrated Data Infrastructure: Linking data for better science and policy. 16th October 2019.
[43] Data in the IDI. (2019). Statistics New Zealand. Wellington, New Zealand.
[44] Integrated Data Infrastructure (IDI) Refresh: Linking Project Summary. (2017). Statistics New Zealand. Wellington, New Zealand.

Acceptability of ADS in New Zealand and Manitoba as International Examples

4.10 Multiple studies have tested public views in New Zealand regarding the acceptability of the IDI. A report for StatsNZ found that the level of public acceptability was rooted in individual experiences and views on the use of data within government and gauged on criteria including the need for the data (proportionality), how the data were going to be used, and by whom, and what protections were in place (for confidentiality, data misuse and misrepresentation).[45] Gulliver (Gulliver *et al*, 2018) conducted qualitative work to understand the conditions under which the IDI could gain social licence (a concept considered in **Chapter 3**). The findings emphasise the need for transparency and public awareness. Those consulted in the study suggested the need for rigorous oversight, that there needed to be a meaningful purpose resulting in public benefits and that mitigations were in place to stop unfair outcomes (which included direct commercial gain, taking advantage or profiling/stereotyping the vulnerable or members of sub-groups).These findings echo many of the issues identified internationally when considering public views on the use of personal health data for research purposes (Aitken *et al*, 2006). It is of interest how data use, and particularly the establishment of data resources and infrastructures, is perceived by cultural groups, particularly in relation to nationhood and autonomy and aspirations. There is some evidence that community ownership of data assets can be empowering and used to help further community wellbeing and autonomy within distinct groups (Boyd *et al*, 2019). This is reflected in the emergence of 'data sovereignty' movements such as those found in First Nation communities in Canada (McMahon *et al*, 2015) and *Te Mana Raraunga* (the Māori Data Sovereignty Network) in New Zealand. **It can be speculated that a devolved and regional perspective on a UK ADS might be perceived as being of more relevance, enabling greater benefit and thus more likely to be acceptable.**

How Would an ADS Operate in the UK?

4.11 The UK does not have a tradition of operating whole population registers that span departments or devolved authorities, either to support government operations or research. The UK has only operated mandatory national population registers, with a citizen ID, in wartime. The last of these was created through the 1939 National Register, which was a population snapshot taken shortly after the outbreak of the Second World War and used for wartime planning and rationing. Following the end of post-war rationing, the register became increasingly contentious, was seen as not being proportionate within some parts of government and was abolished following a court case in 1952. Commentators on this system of government point to concerns about invasion of privacy, the potential loss of civil-liberties (Manton, 2019) and that this way of working does not fit with British tradition, which tended towards localised and siloed 'partial registers' based on individual areas of government function (Agar, 2013). Subsequent interest in developing Population Registers for government administration has not come to fruition (the most recent UK political initiative to launch a new national register and ID card scheme was cancelled in 2010). Resistance to these initiatives appears to result from Population Registers being identifiable, that they are kept up-to-date to reflect changes (as distinct from the decennial census), and that they draw data from across different departments. These features run counter to a belief from some, based on concerns regarding liberty and privacy, that information provided to a

[45] Opus International Consultants. (2015). Public Attitudes to Data Integration.

department through service interaction should be proportionate to the service provision and that access to that data should be restricted to those directly involved.

4.12 It is notable however, that there are examples of data resources in the UK which seem publicly acceptable and that have some similarities to the integrated ADS model, but all have important distinguishing safeguards from the theoretical ADS model:

- The UK decennial Census of Population constructs a cross-sectional Population Register and requires the public to provide information on their health and social status. This has broad public acceptability,[46] although testing of particular questions reveals complex patterns of views on what should be asked to whom;[47]
- Anonymised extracts of the UK Census of Populations are made available for longitudinal research in the form of the Census Longitudinal Studies (Census LSs) and can be enhanced through linkage to routine records (Dibben *et al*, 2017). These operate within each of the UK nations, but can be used in conjunction at a meta-analysis level through their shared designs;[48]
- The Secure Anonymised Infrastructure for Linkage (SAIL) draws together diverse, anonymous individual-level data from the Welsh population. The SAIL databank holds de-identified attribute data and all identifiers are held by the Welsh NHS who act as a linkage TTP. The system was designed following extensive consultation with data owners and Welsh Government, which helped drive acceptability of the resource (Ford *et al*, 2009). The public are involved in SAIL operations through the Consumer Panel for Data Linkage Research;
- Public Health Scotland's electronic Data Research and Innovation Service (eDRIS) provides a liaison service for access to data in the secure Scottish research environment (the National Safe Haven) and supports Scottish Government research and academic research coordinated through the HDRUK and ADRUK networks. Within this, the Scottish NHS Patient Register enables the linkage of data in a consistent manner across health and administrative sources (access to this is restricted to NHS and NRS staff);
- There are now many volunteer population registers for research, such as the Scottish Health Research Register & biobank (SHARE) programme for trial participation in Scotland (McKinstry *et al*, 2017). These demonstrate viability for some purposes - such as trial recruitment - although sign up is likely to be heavily socially patterned. The utility for this in LPS is seen through the NIHR Bioresources: a panel of volunteers who are approached to take part in new studies, or form health volunteer samples for disease specific LPS (e.g., for the Genetic Links to Anxiety & Depression cohort, GLAD). Given the sample size, and selection bias, these are unlikely to align with LPS seeking to enrol representative samples (although there will undoubtedly be some overlap).

[46] The Census and Future Provision of Population Statistics in England and Wales: Public attitudes to the use of personal data for official statistics. (2014). Office for National Statistics. Southampton, UK.
[47] 2021 Census topic research update: December 2018. (2018). Office for National Statistics. Southampton, UK.
[48] Young H. Technical Working Paper: Guide to parallel and combined analysis of the ONS LS, SLS and NILS, July 2009.

4.13 The ONS Secure Research Service in England provides a secure research environment for the analysis of de-identified linked routine records sourced from across different government departments. This is now being expanded via the **ONS Reference Data Management Framework** strategy, which envisages a de-identified 'population spine' to link data from across departments and make these data available for official statistics and research within the ONS secure systems[49]. Whilst in its early stages of development, there are some indications that this resource may fulfil some of the functionality envisaged for the ADS in the LS Review: yet, it is not clear how the LPS community could use any resulting system for sampling, recruitment and follow-up and what operational boundaries would be placed on this use (e.g., if opt-out fieldworker contact protocols would be permitted).

4.14 It is notable that these resources, with the exception of the Census Programme operate on an anonymised basis, their acceptability stemming from the use of rigorous safeguards, transparent operations and particularly anonymous data use that is separated from operational and identity management functions. Within the devolved nations the resources operate in such a way that they are responsive to government questions and can be seen to drive improvements in government function. In this regard, their acceptability may be linked to concepts of 'data sovereignty', although this study has not seen sufficient evidence to be conclusive on this point.

The Feasibility of an ADS for Research

4.14 Information gathering and expert interviews have identified that the ADS model - a centralised, identifiable, whole population research register drawing data from across government departments - as a means to address the problems identified by the LS Review is technically feasible (although likely to be very expensive) but neither proportionate nor acceptable. Whilst the Covid-19 pandemic has highlighted the need for a more robust and response research infrastructure, it has not reinforced a need for an ADS way of working. In relation to options for an identifiable register drawing information from multiple departmental databases, **some experts considered that the negative consequences for confidentiality and the cost of such an exercise would not be justified by the gains to the public good.** It was suggested that anticipated public distrust of, and hostility to, an ADS may impact on public trust in other data-intensive government activities and also that a full population register, without screening to exclude high profile individuals, may pose a risk to National Security. These findings informed the shaping and focus of this study, with the emphasis in the remainder of this report moving to considerations of the existing population registers established within single departments which are critical for government operations and which are being successfully used for research in England (e.g., the NHS Patient Register, the National Pupil Database), in addition to the anonymised research resources described above. The following chapters considers whether these resources, either alone or in combination, could address some or all of the challenges in ways which are feasible, proportionate and acceptable.

4.15 One proposed use for the ADS is to enhance transparency and consent through providing the framework for a single citizen 'portal' where individuals' can access information describing how their data are being used and to set consent indicators to control this use.

[49] Developing an ONS Population Spine. (2019). Office for National Statistics. Southampton. UK.

This functionality is feasible (NHS Digital already 'flag' study membership - for many, but not all LPS - in databases linked to the patient register) but would be technically demanding and costly as it would involve constructing standardised flows of data from across many sources. This study also heard evidence that such a portal can generate substantial concerns amongst the public as being included in a particular study can lead to fears around their health and/or social status (e.g., being selected as a control into a cancer research study could generate confusion over an individual's cancer status) or their service provision. Whilst this concept has merit and warrants further consideration, this functionality is not considered any further in this report.

Key Learning & Recommendations

(1) **An ADS could deliver meaningful benefits to the longitudinal research community:** A population research register augmented with at least minimum attribute information could deliver potentially meaningful scientific and efficiency benefits. It is technically possible (as evidenced by the ONS Digital Census Programme, and as seen in international examples), but would be very expensive to maintain unless it could be repurposed from an existing source;

(2) **The ADS model is not likely to be acceptable nor seen as proportionate to the identified needs:** There is an established public and political antipathy in the UK to the use of *identifiable registers* to support pan-governmental administration, yet a broad acceptance of operational registers within government departments. This antipathy could extend to an ADS for research, which may consequently lack the 'social licence' needed to operate;

(3) **There are existing data infrastructures in the UK which offer some of the benefits of an ADS, but use designs that differ from the ADS model in important ways in order to secure acceptability:** Any population data resource used to help ensure inclusive longitudinal research will need to meet public expectations in regard to privacy controls, information security, transparency and have sufficient public oversight to ensure that data use minimises the risk of harm to individuals and groups and results in public benefits.

Recommendations

(1) A distinct, centralised and identifiable ADS is found to be neither proportionate or acceptable, and this appears a firm position in the evidence from experts. **The progression of the ADS model cannot be recommended at present**, although the current widespread public concerns raised through the Covid-19 pandemic may enhance the legitimacy of new infrastructure or ways-of-working with population data that achieve some of the benefits of the ADS, but stop short of implementing the full model;

(2) It would be **beneficial to monitor the development of the ONS Population Spine** and progress and learning points from the international exemplars of ADS ways of working, described in this chapter, in order to lobby for functionality to support LPS and to help inform the continuing evolution of practice in the UK's alternative approaches.

Chapter 5: The Data Landscape

5.1 Data volumes have grown exponentially in recent years. Lack of data is not the defining challenge relating to using population data within LPS. Rather the challenges relate to 'discovering' which data exist that are relevant to the study objectives, interpreting whether they are fit for purpose, and then, establishing mechanisms to access and utilise them.

5.2 This chapter summarises which data are able to address our identified challenges. Particularly to understand which datasets can act as a 'Population Register' with sufficient scale and coverage to draw inclusive general population or targeted samples of individuals across the UK (i.e., fit for sampling a new cohort, or those with sufficient coverage to link to the diverse LPS across the UK); and then which data exist relating to vulnerable and marginalised populations. In both cases, it is necessary to understand the extent of sub-group exclusions and coverage within each dataset.

Current Population Registers in the UK

5.3 There is no full comprehensive pan-UK population register of all individuals resident in the UK. The closest approximations to this are found in population statistics and the health care registration systems. Both population statistics and health care are devolved matters which results in each UK nation having responsibility for the methods, compilation and management of their relevant data and for the legal basis for which they may be used. There are however, centralised registers of properties for the UK.

5.4 Given that the idea of a centralised and integrated ADS has been discounted, it is worth considering the key attributes of a 'Population Register' for sampling and recruitment and how these differ from the ADS. Crucially, any register must still have as complete coverage of the UK resident population as is possible, it should hold the contact information of individuals and/or households of interest and sufficient attribute data to determine eligibility and to inform stratified sampling and over-sampling approaches (attributes recording status at an individual not area level). Given the sensitivities relating to the ADS, it is not likely that an identifiable and centralised database comprising multiple, linked, data sources would be seen as proportionate or acceptable, and nor will it be likely that records can be transferred across devolved boundaries prior to enrolment and consent. The options are therefore restricted to using a single dataset (at a UK nation level) with options for enhancements using low-sensitivity public domain data (e.g., neighbourhood aggregate indicators), or potentially the flow of individual level data where suitable safeguards and a legal basis can be identified.

5.5 The 'Suitability' of a data source for a sampling frame can be defined as a function of its completeness (coverage and representativeness of the target population); its heterogeneity (inclusion of vulnerable and marginalised sub-groups); its timeliness (specifically in reference to its ability to accurately inform sampling and recruitment); its accessibility; and, whether it is or can made to be 'identifiable'. Here the onus is not necessarily on whether it contains identifiers and contact details, but also whether it can be leveraged for recruitment or tracing and whether linkage to LPS participants can be achieved.

5.6 The primary source of evidence in England & Wales to assess this is the work of the ONS 2021 census programme. The programme has developed a centralised 'Statistical Population Dataset', effectively, a statistical population 'register', which aims to deliver accurate population estimates. The programme has assessed what discrete datasets are available and has compiled the dataset through linking the NHS Patient Register and more lately the NHS Patient Demographic Service (PDS); the DWP Customer Information System (CIS); and data from the National Pupil Database (NPD) and Higher Education Statistics Agency (HESA). This has produced a dataset with high coverage levels, the quality of which has been assessed using the 2011 census. The Electoral Commission has also evaluated the available population registers as part of their program to assess the coverage of the Electoral Roll. Based on these two systematic assessments, the most suitable options for use as a population register to inform LPS sampling and recruitment are: the UK's Census of Population, the NHS Patient Registers, the Education pupil registers and geographical property resources. The suitability of these are summarised in turn below.

The UK's Census of Population

5.7 The closest approximation to a complete population 'register' in the UK is the decennial census taken by the ONS in England and Wales, National Records Scotland (NRS) and NISRA in Northern Ireland. In broad terms the UK Census is *somewhat* suitable for sampling and recruitment. It has excellent coverage (a target of 97.7% population coverage) and heterogeneity, although some sub-groups are under-recorded and missingness is patterned by protected characteristics. While timeliness is currently problematic, given the accuracy of the information erodes over a decennial cycle, the post-2021 digital census is intended to resolve this. The evidence gathered by this study suggests it is currently inaccessible at an *identifiable* individual level and therefore cannot be used to select and contact a sample; although the ability to share anonymised census individual-level data may provide a precedent for sharing data within privacy preserving frameworks. Its primary use within sampling, recruitment and in addressing challenges relating to attrition are in establishing aggregate indicators which can be used to characterise areal units, including for some categories of vulnerability. These have been used extensively in LPS probability sampling and for benchmarking. It also populates the Census LSs across the UK, which are an important LPS resource. Through the use of functionally anonymous research platforms, the Census LSs could contribute towards defining the 'LPS Universe' (see **Chapter 7**) and allow quality assessments where overlap between Census LSs membership and LPS membership exists.

The UK Health Services' Patient Registers

5.8 The NHS maintains three separate register systems: one for England, Wales and the Isle of Man; one for Scotland; and one for NI. Each contains records of all patients registered for care, their NHS ID numbers, their names, addresses, dates of birth, gender and General Practitioner registration details. All three systems are considered to have very high levels of population coverage; and have high temporal coverage given their origins in the Second World War population registration and rationing system. Yet, they are known to have historical data quality issues and identifying or accessing information relating to some vulnerable and marginalised groups is challenging (Boyd *et al*, 2018). Furthermore,

identifying cases across the UK is complicated through internal migration within the UK. **In broad terms the NHS Registers are *highly suitable* as population registers given their high levels of population completeness and that, despite the distributed structure across the UK nations, are accessible and routinely used for sampling and recruiting to research.** The datasets have high levels of heterogeneity although access to some vulnerable and marginalised groups data is restricted, and a sizeable number of patients (between 2-3% of those registered) have opted-out of their data being used for non-consented research. This number has further increased in response to the GPDPR centralised GP records database programme.

5.9 The patient registers contain, or provides routes to, some key demographic indicators used in stratified sampling (age, gender, ethnicity) but will typically have limited information on non-health information such as socio-economic status. The distribution and quality of key indicators is complex: for example, ethnicity is ~95% present when sourced from both primary and secondary sources, but the accuracy of these data are not clear (Wood *et al*, 2021). This could be addressed through linkage to additional data (e.g., birth registration data or neighbourhood indicators). The data are timely given they capture changes of information at all patient interactions, yet this means some groups (i.e., those not needing health care services) are likely to have out-of-date information. The timeliness of centrally recorded health status (e.g., pregnancy) has been problematic for studies seeking to prospectively recruit pregnant women into a birth cohort study. Although, timeliness is likely to improve as midwifery and antenatal service providers (e.g., scan clinics) and software management systems move to interoperable and potentially centralised and queryable databases.[50] Enabling legislation and established ethico-legal routes enable the use of these records for sampling and recruitment: however, variation across the UK may impact on sampling and recruitment protocols (e.g., in NI this study only found precedents for releasing residence information for recruitment, not the personal identifiers of selected individuals). NHS Digital and HDRUK are developing the 'NHS Digitrials' infrastructure to support complex case selection and recruitment: theoretically, this mechanism could be extended to LPS sampling and recruitment (see **Chapter 7**). The registers contain a broad range of contact details, although to date there are only precedents for using postal contacts: which could be challenging where it may be preferable to administer fair-processing information remotely rather than in person.

Civil Registers

5.10 Civil registration of life events (births, deaths, marriages and civil partnerships) is required by law in the UK. The collection and maintenance of the registers is devolved in Scotland to National Records Scotland and in NI to Northern Ireland Statistics and Research Agency. It is managed by the ONS in England and Wales. **The birth registers are *highly suitable* for recruiting a new birth cohort study given their high levels of completeness and relatively timely availability.** Their ability to provide full coverage for timely recruitment may be limited given the mandatory registration period (up to 42 days following birth) can mean that families suffering perinatal and neonatal deaths may be under-represented, as may families where the child is taken into care, those who migrate out

[50] The NHS's Maternity Transformation Programme aims to make wide ranging changes to maternity care provision, including interoperable and accessible digital technologies.

of the catchment area or who become untraceable. Some of the information held within the Civil Registers is effectively in the public domain, given that Birth, Death and Marriage/Civil Partnership certificates are openly accessible to public view. However, additional information collected at birth registration is not made public. New precedents may need to be set to use these data on an opt-out recruitment approach.

Education Census and Attainment Records

5.11 It is a legal requirement that all children receive education in the UK and most will do so through state-maintained education. Across the UK, pupil census and attainment records are collated with national datasets. **In broad terms, education records are *somewhat* suitable for sampling and recruitment. These are the definitive source of linked routine information on education provision and attainment; they also contain important socio-demographic, economic and health indicators; and some data on child behaviours (e.g., absences and exclusions).** Through linkage to further (Individual Learner Record) and higher education (Higher Education Statistics Authority) datasets extends this value, as do sub-sample datasets such as the Child Looked After returns. These data have excellent coverage and heterogeneity of **children in state-maintained education**, meaning that non-maintained early years provision, privately educated and home-schooled children are notable exclusions. Some communities (e.g., Roma, Gypsy, Traveller communities) are at risk of being missing from the data. Timeliness is maintained through semester-based census taking and the key-stage assessments (although the frequency of these have reduced over time). It is accessible at an identifiable individual level, both in terms of attribute data and pupil identifiers. There is the potential for the National Pupil Database (NPD) to inform cross-departmental assessments through the Longitudinal Educational Outcomes dataset**.** NPD has been used to select LPS samples (e.g., the second cohort of the Longitudinal Study of Young People in England), numerous LPS have linked to NPD records (e.g., ALSPAC, MCS, UKHLS) and it has been used to trace participants (e.g., ALSPAC, MCS).

Geographical Data

5.12 All locations in Great Britain are mapped and set to the Ordnance Survey National Grid. Aligned with this, every property in the UK is identified, mapped, given a unique ID number (Unique Property Reference Number, UPRN) and Grid Reference (at 1-meter resolution). Properties are then allocated to a range of official geographies (defined areas of land whose boundaries are known and can be mapped). These can range from very localised geographies (e.g., a Postcode, which contains on average 15 properties) to higher-level health, administrative, political, census, postal and other geographies. All properties can be allocated to higher level geographies by Grid Reference, although at an aggregate level it is typical for mappings to occur using Postcode or the ONS statistical Output Areas.[51] Building a high-quality address register may be critical to study sampling and tracing strategies – with high coverage needed to ensure inclusion and low rates of duplication needed to reduce inefficiencies and avoid participant burden. Public domain information about the natural environment (e.g., pollution estimates, meteorology) and built environment

[51] Office for National Statistics. A Guide to ONS Geography Postcode Products (2016).

(e.g., crime rates, neighbourhood satisfaction survey scores, the availability of services such as parks or GPs within a defined radius of a residence) can typically be linked using either full address or postcode and processed using Geographical Information System approaches. **These geographical data resources are *highly suited* for sampling and recruitment purposes: Address Base Plus is the definitive source of properties in the GB with Pointer providing data from NI. These allow characterisation of property type and assessment of communal establishments.** These property registers have very high levels of completeness, although some characterisations have high levels of missingness, and good heterogeneity (e.g., residences such as caravan parks and some houseboats are also included). The data are maintained on a continuous basis and regular updates are provided meaning the data are timely and have coverage across the UK. They are freely available and are of low sensitivity. LPS have used these resources to inform sample selection in diverse ways. For example, the ALSPAC cohort study defined its eligible catchment area using health geographies from a predecessor to the National Statistics Postcode Directory dataset (Boyd *et al*, 2013).

5.13 **It may be possible to identify vulnerable groups through sub-group specific postcodes,** for example, the authorised sites for Gypsies / Travellers (local authority and private) are assigned unique postcodes (Aspinall, 2014)**. It may also be possible to help tune recruitment strategies using information from these resources,** either directly (e.g., Address Base Plus is able to distinguish communal properties such as care homes which may need different contact/follow-up strategies) or indirectly (e.g., linkage to other neighbourhood information, such as household internet access rates).

5.14 The UK Health & Safety Executive maintain the National Population Database which provides a means to *estimate* the population of the UK at any time through integrating geographical resources (UPRN) with area-based census data. This has been used to estimate the population (workforce) denominator in Covid-19 planning (Chen *et al*, 2021) and may have some potential to inform sampling strategies (although not sampling itself).

Private Commercial Datasets

5.15 It is important to consider that industry will possess customer (population) databases with high-levels of coverage. Examples of this include e-commerce sites such as Amazon (which is estimated to have >80% of active consumers as registered customers), mobile phone providers, high-street banks and building societies and social-media platforms. It is also important to consider that for some marginalised groups, such as the homeless population, these data sources may have high coverage (e.g., >90% of homeless are thought to own a mobile phone; Lemos and Frankenberg, 2015) whereas they may be 'invisible' in many governmental databases.[52] Some of these databases are being actively considered by LPS in terms of sources of 'novel' data collection via record linkage. The acceptability of these approaches is currently being investigated by the ALSPAC cohort (Boyd *et al*, 2019). However, it is not considered feasible that these resources will be

[52] Coverage should not be inferred to mean the presence of an identifiable record in a database, for example, many homeless are likely to be excluded from mainstream mobile phone contracts given their lack of permanent address and credit ratings: they are likely to remain 'invisible', in this instance through the use of effectively anonymous and essentially disposable, pay-as-you-go SIM cards rather than formalised phone contracts.

available to LPS in terms of sampling and tracing for long-term follow-up in the near term, and these sources are therefore out of scope for this study.

Key Learning & Recommendations

(1) There is no centralised population register with complete coverage across the UK. The closest approximation to this is the work being conducted by the ONS to develop a 'Statistical Population Dataset'. Even with unprecedented access to identifiable records from across government departments the trial dataset is still subject to over- and under-coverage of some population groups.

(2) The most suitable datasets for sampling and recruiting participants to LPS in the UK are the national NHS Patient Registers. These contain sufficient identifiers and contact details, but lack socio-economic status information which has previously been used to stratify sample selections by some LPS. There are well established access routes and precedents for this purpose in England; and while there are some precedents in Scotland and NI there is greater uncertainty regarding the acceptability of this in the devolved nations.

(3) There are comprehensive and well-aligned education census returns that cover the UK. These exclude or under-represent some population groups, but could be used for recruiting an accelerated cohort. The ability to access the identifiers from the census returns are likely to vary across the four UK nations.

(4) There are high-quality registers of properties across the UK which are freely available. These can be used as sampling frames in themselves, or to support linkage of area data to other sampling frames.

Recommendations

(1) The most suitable datasets for sampling and recruiting participants to LPS in the UK are the national NHS Patient Registers. These contain sufficient identifiers and contact details, but lack socio-economic status information which has previously been used to stratify sample selections by some LPS. There are well established access routes and precedents for this purpose in England; and while there are some precedents in Scotland and NI there is greater uncertainty regarding the acceptability of this in the devolved nations.

(2) The introduction of the National Opt-Out mechanism has resulted in a large number of individuals opting-out of their data being used for non-consented research. This will effectively block the inclusion of these patients in sampling frames. NHS Digital should be encouraged to assess and report on (at an aggregate level) the health and social patterning of those setting an opt-out flag to inform those drawing inferences from their data.

Chapter 6: Population data approaches to defining 'Vulnerability' and 'Marginalisation' and the implications of these for LPS

6.1 As distinct from legally defined terms such as 'protected characteristics', defining concepts such as 'vulnerability' or 'marginalisation' can be challenging given the different perspectives that will shape answers to questions such as "vulnerable to what?", "marginalised from what or from whom?" and "what are the circumstances shaping the lives of those being considered?" and indeed, whether a person or group would consider themselves to be vulnerable even if labelled as such. The answers to these questions will be temporally specific for those who move in and out of any given category, whether or not they seek support, as a result of changing societal norms, the emergence of new evidence and thinking, and changing political and policy considerations.

6.2 One approach to defining 'vulnerability' within a population is through the use of population data to identify the presence of one or more of an identified range of risk factors and to use this information to categorise individuals as being vulnerable or not (although this approach can overlook that 'risk indicators' can be subjective and overlook important cultural diversity which places different emphasis on mainstream cultural indicators such as educational attendance and attainment). Three prominent examples of this have been found in the UK (which are described in more depth in **Appendix 3**):

- The Children's Commissioner is conducting desk-based research that aims to understand the prevalence of vulnerable children within the population, report to a child vulnerability framework[53] and in conjunction with ADRUK is co-developing the 'Data for Children' initiative[54] to establish a linked population data resource in order to further develop the evidence base relating to vulnerable children.
- The 'Troubled Families' intervention uses a data intensive approach to proactively identify families at risk and to provide these families with suitable support in order to improve their life chances, to improve service provision and to decrease provision costs.
- The Adverse Childhood Experiences (ACEs) framework has emerged as a major theme within epidemiology and public health, as a tool to inform research on the aetiology of health and the social determinants of health (Felitti *et al*, 2019).

6.3 These approaches are hampered as population data (recording service interactions) will record 'late' interventions where children and families have crossed thresholds where they require and become eligible for government intervention. **The Troubled Family and ACEs approaches could therefore miss those who are just below the threshold or who have sub-threshold level risk factors and therefore remain ineligible for services, or those who are marginalised and do not seek help through the standard service providers and may be undetectable in the record.** The advantages of these population data approaches is that they are not impacted by the biases and missingness found in self-

[53] Vulnerable groups and latest data. (2019). Children's Commissioner. UK.
[54] Research initiative harnesses linked government data to improve children's services. (31 July 2019). ADR UK, Swindon, UK.

reported data (although routine records suffer from their own bias), that they can be delivered both at local levels and nationally to suit the needs of different tiers of policy makers and that, in Troubled Families, the weakness of reliance on engagement with service providers is to some extent offset by the involvement of third-sector records.

"There remain significant gaps in data on the support provided to children who do not meet statutory thresholds" (Children's Commissioner 2019).

6.4 The 'breadth' of these population data approaches should be complimented by the 'depth' of LPS and other research studies, which can also help overcome some of the identified weaknesses in using routine records. LPS can provide the early life factors and details of the wider family and social environment that will be missing from population data. A constraint, recognised within the LS Review and by expert contributors to this study, is that **identifying vulnerable individuals within LPS can be challenging, partly as a consequence of vulnerable groups being challenging to recruit and that vulnerability is likely to be associated with the risk of loss of study contact.**

"… alongside that we have shifted… into a kind of thinking about average children and the universal child and we are neglecting the diversity, the heterogeneity and depth of deprivation. …cohort studies do not get at that." (Expert contributor to this study, 2020).

6.5 Vulnerabilities can remain hidden within active LPS participants if there is a reliance on self-reported data compounded by individuals actively concealing behaviours, where reporting is influenced by social desirability bias or where studies are unwilling to ask about sensitive topics or where these are seen as out of the remit of the study. In these circumstances being able to collect data through different modes, or to triangulate data through linkage to routine records, or to seek a different perspective through data collection with independent individuals who know participants (e.g., school teachers, care visitors) may provide new insights and more reliable discovery of a participant's status. Those considering the lived experiences of the most vulnerable have emphasised the importance of 'place' within this and the need to use multi-level analysis to understand how the characteristics of the setting or area in which a vulnerable person lives may influence risk or resilience. The heterogeneity found within LPS sampled from different cities and regions (e.g., those in economically advantaged areas of the UK such as ALSPAC or the Edinburgh Study of Youth Transition and Crime, or those in disadvantaged areas such as Born in Bradford) or LPS with national coverage with sufficient population coverage in distinct cities/regions (e.g., UKHLS) may be well placed to inform this type of analysis.

6.6 This suggests that for LPS to inform policy development and thinking in this area they will need to consider how to accommodate the following considerations in their designs (although, as described previously it is not necessary for each LPS to include all groups and that these considerations need to be made within the context of the aims and design of the study)

- To understand the index participant, it is necessary to understand the others within their household and their wider social network. For example, to understand

vulnerability in children it is necessary to capture information on the other children and adults they live with, or the support networks they may have with other adults;

- There needs to be specific data collected on the identified risk factors for vulnerability and the categories of being vulnerable, and on mitigating factors which may provide resilience to vulnerability;
- The collected data needs to be sufficient to understand the dynamics of vulnerability: to capture data frequently enough to measure drift in and out of being considered vulnerable (i.e., event changes between data collection waves) and to be responsive to changes in the perceived categories of vulnerability over time;
- Arguably the most vulnerable members of society will be outside of family and household structures: therefore sampling and follow-up strategies must consider those in institutions (e.g., children in residential care, the elderly in care homes, prisoners), those who are homeless and marginalised groups with transitory lifestyles;
- The experiences of the vulnerable, or clusters of vulnerabilities, may play out differently, for example according to the characteristics of the location, or between England and the devolved authorities. This suggests a need to characterise the place and to oversample within heterogeneous geographical areas;
- Intersectionality is important. Assessing the relationships between categories of vulnerability will require access to data across broad domains;
- LPS provide an opportunity to collect self-reported adversities which may be unreported to social workers or other care givers and to understand the trajectories of those not seeking help or for those under service-provision thresholds;
- To not only measure adverse experiences, but also to collect data on individual's strengths, assets, their resources and the equivalents relating to members of the household and the wider communities.

6.7 Population data can in theory be helpful in informing LPS sampling strategies, for example through incorporating vulnerability status in stratification or over-sampling strategies. However, this study received consistent evidence that for factors relating to the longevity of the study (i.e., retaining the flexibility to study a wide range of factors) and simplicity of subsequent analysis it would be beneficial to implement a simple sampling strategy which did not over sample or selectively sample by vulnerable groups. This does not preclude the option of conducting aligned studies or sub-studies which focus on specific vulnerable or marginalised groups (such as children taken into care). This view is supported by an accompanying report in the ESRC's Population Laboratory initiative (Sullivan *et al*, 2020).

6.8 They are also likely to be of value in the follow-up of participants exposures/outcomes, for example through providing a linkage resource to inform missing data strategies and bias assessments. These data may also be of use in designing participant engagement and inclusion plans as the insights gained from understanding the presence of distinct participant sub-groups may help tailor contact protocols and to identify and recruit relevant public representatives and/or third sector and community groups.

Key Learning & Recommendations

(1) Defining vulnerable individuals using population data will be challenging given that not all those who are vulnerable will be in receipt of services, that the information needed to make this classification may be inaccessible, that those at risk of being vulnerable may be missed as their vulnerability is temporal, that some individuals 'drift' in and out of being considered vulnerable, that some may be excluded or under-represented due to threshold based definitions and that some labelled as being vulnerable may not consider themselves as such.

(2) The complexity of defining vulnerable groups and the diversity of these should be considered in relation to the general purpose 'resource building' remit of many of the general population UK LPS.

(3) These challenges reinforce the evidence that while the use of population data may help address the challenge, **this is likely to need to be done in conjunction with rigorous fieldwork approaches**.

Recommendations

(1) In light of the recommendations in the sampling report and evidence gathered in this study's expert interviews, **it is recommended that new studies should emphasise sample heterogeneity and general inclusivity rather than complex sample selection based on the inclusion of one of multiple vulnerable groups.**

Chapter 7: New enabling infrastructure and ways of working for inclusive Longitudinal Research

7.1 The LS Review and now this study have identified benefits that could arise through new ways of working with population data to facilitate inclusive longitudinal research. This study has identified a number of key uses of population data for inclusive research which could be facilitated by an ADS and are not currently available:

- To utilise individual level data for sample selection and stratification, and then to contact selected individuals using information held in their official record;
- To assess population and sub-group coverage and inclusion across the sum total of LPS participants in addition to assessing inclusion at a study level;
- To systematically follow-up participant status and to use linked records to facilitate the inclusion of those who are vulnerable and/or marginalised using a centralised resource.

7.2 Given that an ADS is not recommended, this chapter summarises two potential mechanisms to achieve some of these aims which are substantially different from an ADS in key design factors and are based on considerations as to what infrastructure options and ways of working could be seen as proportionate and socially acceptable. **Any action to realise these mechanisms should be made with consultation and involvement with the public and other stakeholders.**

An outline protocol for 'Privacy Preserving Sampling and Recruitment'

Objective, Challenge and Precedent

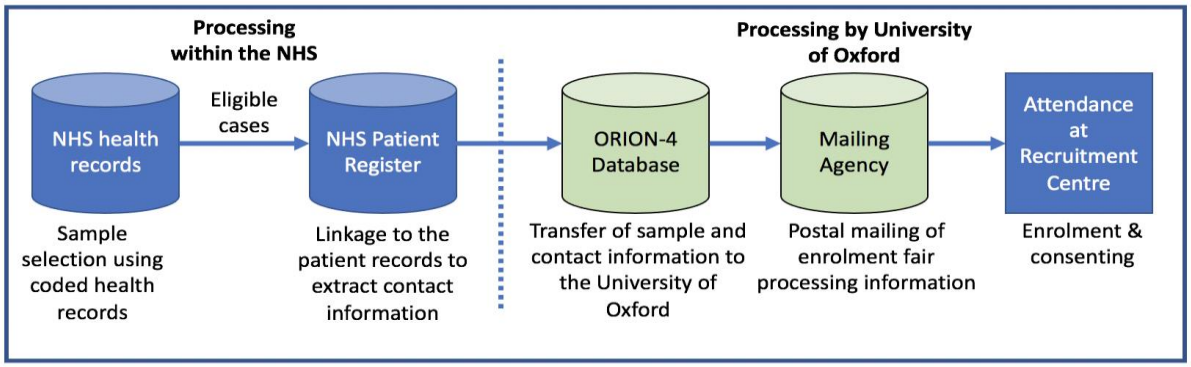
7.3 The objective is to enable the use of individual attributes and contact data from one or more sources for sample selection and recruitment purposes. The ambition here is to make the sharing of data of identifiers acceptable to data owners through promoting a secure and legal route for this and engaging them in furthering the representativeness and equity of longitudinal research. There would be strong advantages in a 'live' system which could support the dynamic assessment of recruitment performance.

7.4 The challenge is to overcome any data owner reticence for this use of their data which stems from it being identifiable, to demonstrate that this use is legal, ethical and that it includes sufficient design characteristics so that it is likely to be considered acceptable to a fair and reasonable member of the public. That there is a specific challenge to sharing identifiable data across departments and to overcoming the 'consent for consent' paradox.

7.5 The most useful precedent for this is the NHS DigiTrials system (see also 2.29 - 2.30) where the NHS conducted sample selection using coded health outcome records, linked selected information to the NHS patient records and provided the personal identifiers to the University of Oxford to conduct a recruitment mailing campaign (**Figure 1**). The flow of information in this way uses s.251 Support to set aside Duty of Confidentiality in England and Wales, and approval from PBPP in Scotland. Note that this does not provide a

precedent for 'consent for consent' as the recruitment asked individuals to attend a study centre to enrol (i.e., response was an active decision of the individual).

Figure 1: illustrative flow diagram for the NHS DigiTrials and ORION-4 trial.



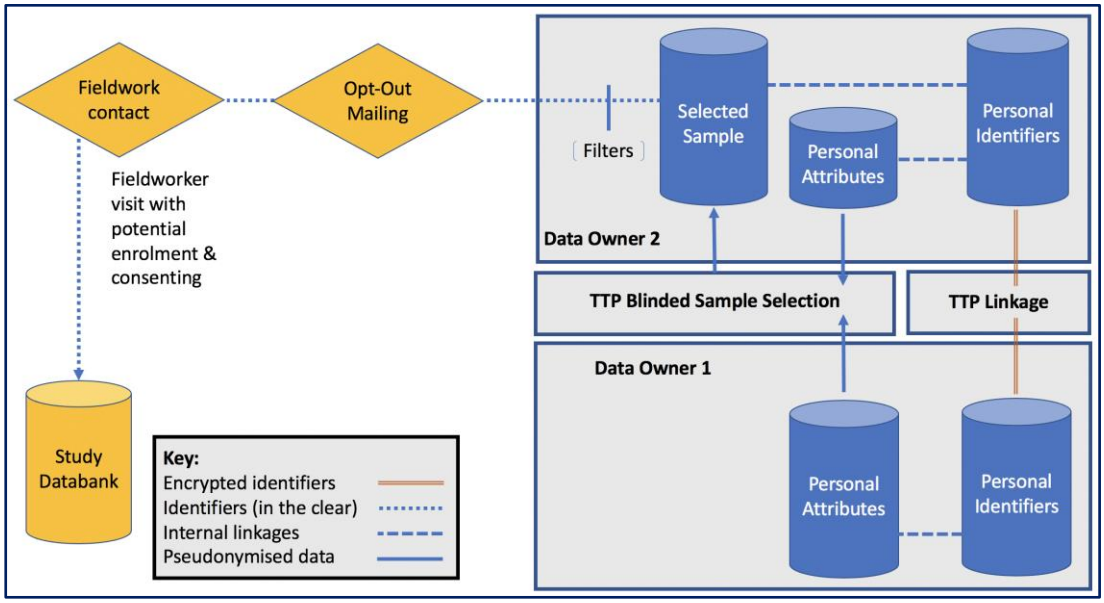
Outline solution

7.6 This outline protocol building on the NHS DigiTrials framework to illustrate how data can be provided across departmental databases to inform recruitment. A *theoretical example* would be to use the national Troubled Families database to select individuals with a range of adversities and to link them to contact information held within the NHS patient register. In this example membership of the Troubled Families population would be considered sensitive (given it indicates a potentially stigmatising status), the information within the database would be considered highly sensitive (given it details specific adversities) and the transfer of this data to the NHS for the purposes of supporting research would be considered – without consent - to breach expectations of confidentiality and infringe on the right to privacy.

7.7 The privacy preserving protocol (**Figure 2**) would address this breach of confidentiality by using privacy preserving record linkage to encrypt the flow of identifiers and for the attribute information to be masked to all except the original data owner. The disclosure from this system would be identifiable contact information (name and address) only. Although, this could be supplemented along with non-disclosive indicators (indicating priority recruitment individuals) so that fieldworkers could target resources to harder to reach individuals. The initial mailing to individuals (providing a means to opt-out) could be conducted by the data owner, by a TTP or by the study team (as seen in the NHS Digitrials example) with a secondary - and ideally in person - contact being made by the study team or fieldwork agency.

7.8 This protocol mitigates the concerns raised by an ADS through not seeking to generate a whole population register, managing the exchange of data between departmental databases in a privacy preserving manner (rather than seeking to integrate data) and through ensuring the users of the contact information are blinded to any source attribute data and indeed, the users could even be blinded to the source of the data at the point of recruitment (although this would eventually become known as the methodology was reported). Careful consideration would need to be made as to whether this blinding would be considered reassuring (in that the approach was made on the basis of the provision of a carefully selected sample designed to facilitate inclusive research where the provenance of the information was restricted to only those who need to know it) or whether this lack of transparency would be considered unethical and risking social licence.

Figure 2: illustrative flow diagram for multi-party privacy preserving sampling and recruitment



The protocol sends fair processing materials to help set a reasonable expectation with individuals as to how their data are being used and to provide an opt-out mechanism implemented prior to the release of data from the co-ordinating data owner to those conducting the recruitment fieldwork. The 'consent for consent' paradox is addressed both through these measures and through implementing safeguards to control risks of harm and implementing rigorous oversight mechanisms. Where NHS records are used in England, this would have to demonstrate potential benefits to the health and social care system and should also respect National Opt-Out objections unless there is particularly strong evidence that the patterns in these would bias the recruitment to the detriment of those who had set these objections (public dialogue would be needed to explore this issue).

What Next?

7.9 **This is an outline solution for a potential new way of working, and any attempt to realise it would require greater investigation and consultation.** However the following may create challenges in doing this:

1) currently the data science community and those tasked with sharing data are at full capacity dealing with Covid-19 related issues and this may be a barrier to new system development for an imminent new study;

2) the cost of developing this system may be out of proportion to the benefits to sampling and would likely only be justified if it were sustainable, interdisciplinary and generalisable to a wider range of applications (e.g., an enhancement to NHS DigiTrials rather than a bespoke solution for longitudinal studies);

3) the solution mitigates the risk of harms relating to 'consent for consent' rather than providing a fully risk free solution: this could still generate risk aversion from data owners; still raise concerns over 'consent for consent' and the legitimacy of releasing

named contact details at any stage of the process; and does not remove the need to make substantial efforts to communicate the purpose and safeguards of this to the public (as identified in the public dialogue work). This approach is aligned with observations that data science infrastructure should be based on citizens' trust and a social licence and not a technical solution alone (e.g., Jones and Ford, 2018; Moore *et al*, 2016).

7.10 Since the evidence gathering for this report, the National Core Studies for Covid-19 programme - which involves the four nation NHS agencies providing data for research and the ONS - have developed data sharing mechanisms to create the 'ONS and NHS Digital joint health data asset'. This brings together health and social records within the ONS Secure Research environment. It remains to be seen if these will be sustained and what the permitted scope of use cases of these data are. If this data sharing was sustained into the long-term then this could reduce the need for privacy preserving mechanisms within this overall protocol. Those designing a new cohort should explore the options for leveraging this dataset before considering investing in this protocol.

An outline protocol for a centralised LPS linkage infrastructure for the assessment of inclusivity and follow-up through linkage to population data.

Objective, Challenge and Precedent

7.11 The first objective would be the creation of a resource which can co-locate the study data of many (ideally all) LPS with linked health, social and environmental records. The resource would need UK coverage and to include LPS from biomedical and social science disciplines. This could form a de-identified 'Participant Spine': determining the 'UK LPS Universe' of (ideally all) participants. This could then be assessed in terms of coverage and inclusivity against whole population records. Taken as a whole – and with sufficient LPS buy-in – this assessment could form a mapping of the coverage of LPS and inform assessments of gaps in coverage and the development of future studies and engagement strategies. There would need to be a sufficient 'air gap' between the studies (who maintain both participant identifiers and attribute data) and the resource/resource management in order to ensure this new infrastructure was functionally anonymous.

7.12 This resource would form the framework for a novel infrastructure, which could support efficient linkage informed follow-up through managing centralised governance, contractual, data flows whilst minimising direct costs (data charges, infrastructure overheads). For this, the objective would be to develop and maintain a sustainable platform for researchers and those developing highly curated value-added data resources (such as tightly harmonised and integrated datasets around health or social themes). Whilst it would need to curate the data it holds, the processing of these should be minimised so the infrastructure has maximum flexibility to support diverse projects and approaches. It should be transparent in its operation, auditable, and capable of sharing the knowledge and research tools (e.g., code definitions, processing and data derivation syntax) developed in projects to new users.

7.13 The primary novel challenge is for this centralised resource to be an effective 'broker' of the respective needs of the contributing studies (and through them, their participants), the

contributing data owners and other components of the UK data science community. This brokerage would relate to data, access and legal requirements:

- It will need to manage the numerous and diverse assurances that studies have made to participants regarding how their data are used: these form part of the 'social contract' between participants and the study; and legally help set participant expectations as to how their data will be used. The critical 'control points' traditionally implemented by the studies should stay within the studies control (i.e., determining which linkages are established, which data are provided and how these are de-identified, which participants are included, which users can use what data for what purpose);
- As an infrastructure to support linkage informed research, it will need to broker access to the data. While access review should be retained by the studies where this is an expectation of participants, the resource should negotiate delegated control of the review process on behalf of data owners (against an agreed protocol) to aid responsiveness;
- The resource should negotiate centralised flow of data sharing agreements and a common legal basis for this (likely with variations for devolved nations). This will remove an administrative burden from the studies and minimise the equivalent burden on the NHS and administrative providers.

This brokerage function is a novel and non-trivial task and will require those developing such a resource to work across the LPS community to develop a common 'level playing field' of governance standards and participant expectations. Failure to address this effectively could jeopardize the study-participant trust relationship and the confidence of data owners which supports the flow of data. The funders would need to consider how to incentivise studies to take part and to support the resource implications of doing so.

7.14 The infrastructure would also need to meet the requirements of multiple NHS and other data owners, to identify an ethico-legal basis and to accommodate the differences in legislation and access procedures found across the UK. It is likely this will require the use of a Trusted Research Environment (TRE)[55] approach coupled with the development of information security management system and a governance framework which is accredited to leading standards (e.g., ISO27001 and Digital Economy Act certification for the linkage and processing of data).

7.15 There are a number of precedents which are useful when considering models for this resource:

- Many LPS already have set frameworks with their participants for the use of linked health and social records within the study research programme. Whilst any use for such a resource would likely be outside the bounds of what was previously described, a suitably designed infrastructure to retain enough of the existing study control structure to mean that re-consent was not necessary (although additional fair

[55] A Trusted Research Environment is a technical and governance solution for managing the linkage and use of data within a secure environment. It provides access to data managers for data integration and processing and working space for researchers to conduct analysis. It would have a defined access mechanism and its activities would be transparent to the public. See the HDRUK green paper for detail.

processing and opt-out would likely be needed to set 'reasonable expectations' for data flows and use);

- Likewise, most LPS have established principles where de-identified data are shared with external researchers under a set of conditions (typically that the data are de-identified, that the purpose and user have to be approved by the study, and that the objective of the research is to improve the public good);
- Some studies have already implemented designs where data are accessed and analysed via third party infrastructure which is run by the study (e.g., ALSPAC's Data Safe Haven which uses a UK Secure eResearch Platform) or is outsourced. The UK Data Service provides such functionality across ESRC supported studies and is now facilitating the provision of linked study-health records established by the studies. Dementias Platform UK provides an example of this where data are drawn together from many contributing studies and harmonised for dementias research.

Outline Solution

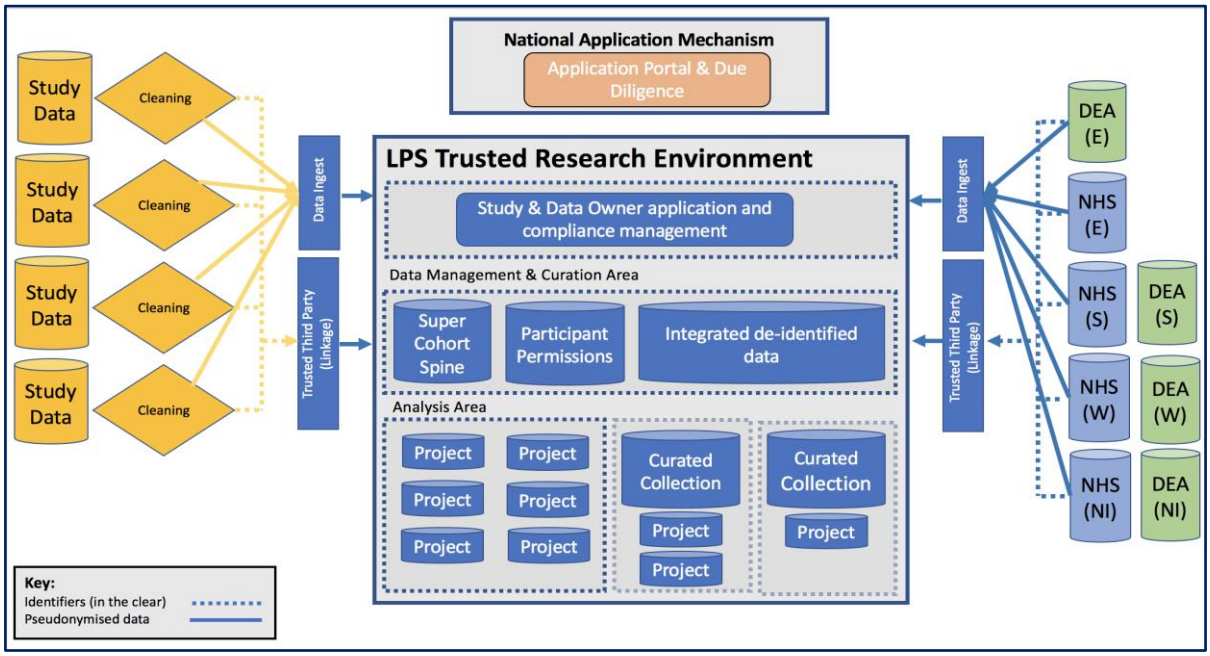
7.16 A new TRE (**Figure 3**) is established as a secure environment in which de-identified data can be ingested, processed and provisioned for research. It would operate as a 'reading library' where all processing and analysis occurs within the infrastructure and all outputs are restricted to population level aggregates and statistical outcomes. It would use existing TRE 'infrastructure as a service' such as that offered by the UK Secure eResearch Platform (UKSeRP). A TTP would be required to manage the flow and processing of identifiers for linkage purposes: a central TTP could act as a linkage 'broker' to disseminate identifiers across the UK NHS data provision agencies and other holders of routine records. The host institution - acting as Data Controller under data protection legislation - would enter into data sharing and processing contracts with studies, the controllers of routine records and with the provider of the TRE services. Applications to use the data would be managed ideally through a centralised national mechanism (e.g., HDRUK Gateway) and would be assessed against an existing framework (e.g., the Five Safe's framework). Applications would be 'triaged' by those operating the infrastructure (to assess feasibility and to conduct due diligence checks), distributed to contributing studies for local approvals, and reviewed centrally to manage the requirements of the NHS and other data owners.

7.17 The insights from this study suggest that the legal basis for such a resource would be:

- For DPA, the infrastructure would operate under 'public interest' and 'scientific research' provisions (Article 6(1)(e) and Article 9(2)(j) subject to Article 89(1));
- To address Duty of Confidentiality, the TRE would only contain functionally anonymous data (where re-identification was not reasonably likely) and where the bulk of processing would therefore not result in a breach of confidence. The flows of identifiers for linkage would be managed by setting a reasonable expectation and either gaining consent or through offering an opt-out and using a mechanism to set aside the requirement for consent where consenting would not be practicable;
- The NHS would be able to flow de-identified data using their statutory legal gateways, although the basis for this in Northern Ireland would need consideration.

Non-health records could potentially flow using NHS legislation where they are 'related' to health status and to be used for purposes designed to inform the health and social care system, or using DEA provisions (subject to the infrastructure gaining relevant certifications).

Figure 3: illustrative schema for the centralised LPS linkage trusted research environment.



7.18 This infrastructure would mitigate some of the concerns raised by an ADS model through only processing de-identified data, being minimised to only include the data of LPS participants and not the wider public, not enabling/requiring the exchange of data between departmental databases, and that all flows and processing of data are governed by contract. It is not intended to be a resource for sampling and recruitment (as it only contains records of existing enrolled participants). It would be necessary for studies to communicate with participants to set reasonable expectations for this new data and to offer an opt-out: mechanisms would need to be implemented to apply changes in participant permissions over time. Critically, the studies would retain the key 'control points' that participants expect them to implement in order to respect their rights. Public/participants would need to be involved in the application process and the shaping of the governance frameworks and safeguards used to minimise risks to confidentiality. The resource would be subject to audit and independently assessed to relevant standards. Its operations, the data it holds and the use of these data should be transparent and externally verifiable.

Forming a de-identified 'participant spine'

7.19 This infrastructure could establish a persistent ID link between the participant record within the TRE and the databases of external data owners (e.g., the NHS patient registers). However, the study received evidence that a sizeable minority of LPS participants take part in multiple studies: which would need addressing within this infrastructure in order to make

accurate coverage inferences and to reduce statistical error (given that many statistical approaches are made on the assumption of independence of samples). Given the lack of a reference population spine that an ADS could have enabled, the challenge in this context would be to establish a de-identified participant spine of unique individuals based on the identifiers provided by the studies. This could be achieved through the TTP using probabilistic linkage techniques to 'de-duplicate' incoming participant identifiers. This would likely work through compiling the incoming information into a single master list and then linking it to a copy of this list. Change over time could be managed by linking new incoming information against de-duplicated output from this process (meaning the TTP would have to retain a copy of these identifiers). Rigorous version control and referencing would be needed. This approach is compatible with current linkage theory but would benefit from theoretical testing using synthetic data to identify whether this would result in inconsistent linkage outcomes over time, and that these may be more likely to occur within vulnerable sub-groups (see the recent ONS series on linkage quality for a discussion on this).

What Next?

7.20 The theoretical considerations of this model from this study have been taken forward by the Longitudinal Health & Wellbeing National Core Study (LHW NCS) and a group of 'vanguard' LPS who have established the UK Longitudinal Linkage Collaboration (UK LLC) as a novel and globally unique TRE for co-location of study data with linked health and non-health administrative records to a similar protocol as summarised above. The UK LLC is being led by the University of Bristol with the University of Edinburgh, is based on UKSeRP infrastructure, and is working with ~20 UK studies representing the four UK nations including the ESRC's major longitudinal investments as well as those primarily supported by MRC and the Wellcome Trust. The UK LLC as currently operated is a resource for Covid-19 research only, although it is designed to be scalable to enable other research purposes. This extension would require modifications (but not fundamental changes) to its contractual agreements, the approval of the contributing studies and would need to be informed by involvement of participants to help ensure the safeguards are appropriate and that this way of working in longitudinal research is acceptable beyond the current circumstances of the Covid-19 pandemic. Within this, the UK LLC is working with University of Swansea (the developers of the UKSeRP) to establish the 'linkage brokerage' functions sufficient to establish a de-identified participant spine and for this to be linked to external data owners.

7.21 Through aggregating samples from contributing studies, the UK LLC could potentially provide viable sample sizes for studying sub-groups and thus enable better representation of under sampled populations or groups underserved by the coverage of existing studies (e.g., addressing the current under-coverage of adolescents in the UK by aggregating samples of 3rd generations of existing studies [such as ALSPAC and the Twins Early Development Study] with adolescents in family and household studies [such as Generation Scotland and UKHLS]).

7.22 The UK LLC has a Public and Patient Involvement and Engagement strategy which will enable the public to play an active role in the resource (e.g., through involvement in the project approval process), to helping develop materials to communicate the design (e.g., a UK LLC animation), purpose and benefits of the resource with the public. The LHW NCS have developed a statement to make clear the boundaries for data use and the safeguards

which have been deployed (see **Panel 8** and the LHW NCS website). The objective of the UK LLC governance framework is for the studies to retain all the key decision-making powers (who uses whose data for which purposes) and thus not erode the participant-study trust relationship.

Panel 8: Longitudinal Health & Wellbeing Statement for our study participants about the use of your data and NHS data for research.

Our commitment to you:

- We never use personal identifiers such as your name or address in any of our research.
- We only use personal identifiers such as your NHS number to link the information you give us to your health records.
- This de-personalised data is used solely for research in a secure, confidential space, called a "Trusted Research Environment". Results cannot leave the environment until an independent checker has confirmed that no individual can be identified.
- Approved researchers can only access the trusted research environment once they and their research question have been checked and approved.
- No data is shared for profit making purposes. We do not sell your data, and we never will. Any researcher using your data signs up to this commitment.
- As a community of studies and data scientists we manage your data ourselves. We do not outsource to private companies.
- Our ways of working are reviewed by independent ethics committees and volunteer study members.

7.21 Through the National Core Studies, the UK LLC – with the Data & Connectivity National Core Study (HDRUK, ADRUK, ONS) - are investigating mechanisms to link non-health administrative records into the resource through DEA provisions. The DEA provides flexibility by allowing the processing of identifiers for matching, linkage and de-identification process ('preparation') to be undertaken by accredited agents and for de-identified data to be stored and made available for analysis ('provision') within accredited secure environments[56]. Meaning that the structure of the linkage and analysis processing is not predetermined and can potentially involve a wide range of existing data owners. Currently, there are accredited processors across the four UK nations: including the UKSeRP/SAIL research databank (as a processor and provider) and the UK Data Archive (as a provider)[57]. This functionality and precedent suggest the viability of including linked non-health records in this way as the linkage could be handled by UKSeRP (processor) and the provision by the UK LLC if it gains accreditation to the provider. While there is no direct precedent for this with longitudinal studies: a precedent for this manner of flowing data has been set in Wales where linked administrative records have flowed into the SAIL Databank under DEA provisions via the NHS Digital Health Care Wales TTP (Bedston *et al*, 2020), which is allowing the records of vulnerable children (children in the family law system) to be used in research.

7.22 This study has gathered consistent evidence about the value of transparency of operations in terms of building public trust coupled with public dialogue evidence reinforcing

[56] UK Statistics Authority Digital Economy Act Processor Accreditation Guidance. 2020.
[57] See ONS site for a list of accredited processors under the Research Strand of the Digital Economy Act.

the value of independent oversight of data science activities. A new model for transparency is being implemented by OpenSAFELY which publishes all queries made by researchers to the research database. This provides a true publicly auditable record of all actions of those using the system. The mechanism has an important secondary benefit of promoting the sharing and recycling of research tools to new projects and users. It might be possible to implement such a framework within the UK LLC TRE. This will be important to consider as part of efforts to ensure the sustainability and acceptability of the infrastructure.

Key Learning & Recommendations

(1) The privacy preserving sampling and recruitment protocol is described here as a theoretical approach which is primarily intended to describe a range of tools that are available to those recruiting to a study and some, all or none of which may facilitate negotiations with data owners when seeking permission for sampling from government department databases and using opt-out based recruitment approaches;

(2) A centralised LPS resource for record linkage provides a means to address uneven across data domains and the high resourcing barrier to entry to establishing and maintaining linkages;

(3) This centralised resource would provide a mechanism to comprehensively and systematically assess population coverage across a range of LPS and thus identify if the 'structural holes in coverage' which the LS Review raised concerns over exist and to inform strategies for addressing these;

(4) An implementation of this concept is now being implemented as the 'UK Longitudinal Linkage Collaboration' as part of the National Core Studies for Covid-19 research programme. The UK LLC collaboration already includes 20 studies with pan-UK coverage and those with biomedical and social science remits.

Recommendations

(1) It is recommended that **a privacy preserving protocol is considered to mitigate perceived privacy risks during sample selection and recruitment** and to improve performance where access to the flow of identifiable data for opt-out recruitment approaches cannot be secured.

(2) It is recommended that the LPS community c**ontinue the development of an interdisciplinary and pan-UK centralised mechanism for cross-cohort investigations utilising population data: the UK LLC is being implemented as such a resource** and the sustainable funding for this (or an alternative solution with similar functionality) should be considered through the PRUK programme.

(3) It is recommended that t**he ESRC consider funding options for a comprehensive assessment of LPS population coverage using the UK LLC and whole population databases.** This should be presented as a competitive call to the longitudinal research community.

Chapter 8: Conclusions and next steps

8.1 The ESRC commissioned this report to explore the use of population data in the UK, with a focus on exploring the options and feasibility for an ADS and to consider the potential for population data to help ensure that longitudinal research is inclusive and representative of the UK population. Hence, the results of this report are of particular interest to researchers and policy makers. Accordingly, ESRC will promote this report to the scientific project teams of current and future cohort studies and to the wider academic, policy and third sector communities to encourage wide-spread use of this report. The report will also help inform ESRC's strategy and future activities in relevant areas.

8.2 The study has concluded that an ADS is neither proportionate nor acceptable. However, detailed consideration of the issues raised in the review has reinforced the importance of statutory requirements, ethical principles and public involvement and engagement as the key drivers to ensure, at a holistic level, the longitudinal research community's research programme is inclusive and makes best endeavours to include harder to reach communities. To help progress the findings of the LS Review, the study has considered the key issues in detailed discussions in the preceding chapters of this report. Each chapter finishes with key learning points and recommendations.

8.5 The National Core Studies for Covid-19 has provided the urgent use case to developing the UK LLC as an example of centralised linkage infrastructure. The development of the model has drawn on this study and the wider ESRC Population Laboratory theme. The means to progress the model, and potentially generalise it to broader purposes, will be again based on this study's insights, specifically those drawn from the public dialogue exercise and the thinking around the benefits that an ADS could have delivered. The long-term sustainability of the UK LLC, or similar ways of working, are currently being assessed during considerations to design a PRUK.

8.6 More broadly, the research response to the Covid-19 pandemic and the public awareness around this, has increased the potential for using population data more widely. However, it is not a given that any new ways of working are sustainable. The LPS community should strive to ensure widespread public awareness and work with public representatives to ensure a framework is in place to make these innovations sustainable and to set a reasonable public expectation for this based on meaningful safeguards and working practices that are sufficient to maintain a social licence for our work.

8.7 The ethos of the UK LPS community is already intrinsically inclusive: yet, more needs to be done at a study and community level to engage the harder-to-reach, to build understanding and trust, and to enable our research to be inclusive across the socio-economic and health spectrum. Adopting the above new ways of working in a sustained way, whilst maintaining a social licence, could bring significant benefits. More representative and inclusive sampling frames, with follow-up through improved linkage, is likely to generate efficiencies and better-quality data. This can enable longitudinal research to better inform the decision processes of policy makers, in turn leading to improved benefits and outcomes for all target populations. Sufficient and sustained resources should be made available to support LPS develop inclusion plans and to help realise these through integrating population data with high quality fieldwork and community relationship building.

References

Aitken M, Jorre JD, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC medical ethics. 2016 Dec;17(1):73.

Aspinall P. Hidden needs. Identifying key vulnerable groups in data collections: vulnerable migrants, gypsies and travellers, homeless people, and sex workers. Kent. 2014.

Azad MB, Konya T, Persaud RR, Guttman DS, Chari RS, Field CJ, Sears MR, Mandhane PJ, Turvey SE, Subbarao P, Becker AB. Impact of maternal intrapartum antibiotics, method of birth and breastfeeding on gut microbiota during the first year of life: a prospective cohort study. BJOG: An International Journal of Obstetrics & Gynaecology. 2016 May;123(6):983-93.

Batty GD, Morton SM, Campbell D, Clark H, Smith GD, Hall M, Macintyre S, Leon DA. The Aberdeen Children of the 1950s cohort study: background, methods and follow-up information on a new resource for the study of life course and intergenerational influences on health. Paediatric and perinatal epidemiology. 2004 May;18(3):221-39.

Bird PK, McEachan RR, Mon-Williams M, Small N, West J, Whincup P, Wright J, Andrews E, Barber SE, Hill LJ, Lennon L. Growing up in Bradford: protocol for the age 7–11 follow up of the Born in Bradford birth cohort. BMC public health. 2019 Dec;19(1):1-2.

Beale N, Peart C, Kay H, Taylor G, Boyd A, Herrick D. 'ALSPAC' infant morbidity and Council Tax Band: doctor consultations are higher in lower bands. European journal of public health. 2010 Aug 1;20(4):403-8

Bécares L, Kapadia D, Nazroo J. Neglect of older ethnic minority people in UK research and policy. BMJ 2020;368:m212.

Bedston S, Pearson R, Jay MA, Broadhurst K, Gilbert R, Wijlaars L. Data Resource: Children and Family Court Advisory and Support Service (Cafcass) public family law administrative records in England. International Journal of Population Data Science. 2020;5(1).

Benzeval, M., Burton, J., Bollinger, CR., Crossley, TF. (2019) Methodological Briefing: The representativeness of Understanding Society. Briefing Colchester: ISER, University of Essex.

Berthoud R, Fumagalli L, Lynn P, Platt L. Design of the Understanding Society ethnic minority boost sample. Colchester: Institute for Social and Economic Research, University of Essex (Understanding Society Working Paper 2009-02). 2009 Dec 1.

Bird PK, McEachan RR, Mon-Williams M, Small N, West J, Whincup P, Wright J, Andrews E, Barber SE, Hill LJ, Lennon L. Growing up in Bradford: protocol for the age 7–11 follow up of the Born in Bradford birth cohort. BMC public health. 2019 Dec;19(1):1-2.

Black LA. Health and Social Care (Control of Data Processing) Bill. Northern Ireland Assembly; 2015 Sep 10. Available from: http://www.niassembly.gov.uk/globalassets/documents/health-2011-2016/legislation/control-of-data-processing/papers-from-department-and-others/9.-research-paper---key-provisions-of-the-control-of-data-processing-bill-10.09.15.pdf

Blackburn RM, Hayward A, Cornes M, McKee M, Lewer D, Whiteford M, Menezes D, Luchenski S, Story A, Denaxas S, Tinelli M. Outcomes of specialist discharge coordination and intermediate care schemes for patients who are homeless: analysis protocol for a population-based historical cohort. BMJ open. 2017 Dec 1;7(12):e019282.

Bliddal M, Liew Z, Pottegård A, Kirkegaard H, Olsen J, Nohr EA. Examining Nonparticipation in the maternal follow-up within the Danish national birth cohort. American journal of epidemiology. 2018 Jul 1;187(7):1511-9.

Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, Brand CA. Data linkage: a powerful research tool with potential problems. BMC health services research. 2010 Dec;10(1):346.

Bonevski B, Randell M, Paul C, Chapman K, Twyman L, Bryant J, Brozek I, Hughes C. Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. BMC medical research methodology. 2014 Dec 1;14(1):42.

Booker CL, Harding S, Benzeval M. A systematic review of the effect of retention methods in population-based cohort studies. BMC public health. 2011 Dec;11(1):1-2.

Boreham R, Boldysevaite D, Killpack C. UKHLS: Wave 1 technical report. London: NatCen. 2012 Jan.

Borkowska, M., (2019) Improving population and sub-group coverage: who is missing and what can be done about it? Understanding Society Methodological Briefing Colchester: ISER, University of Essex.

Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, Molloy L, Ness A, Ring S, Davey Smith G. Cohort profile: the 'children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. International journal of epidemiology. 2013 Feb 1;42(1):111-27.

Boyd A, Thomas R, Macleod J. NHS Numbers and their management systems. London, UK: CLOSER; 2018.

Boyd A, Coleman G, Spence E, Park A, Hardy H. (2019). An outline framework for the efficient onward-sharing of linked Longitudinal Study and NHS Digital records. London, UK: CLOSER, University College London.

Boyd A, Gatewood J, Thorson S, Dye TD. Data Diplomacy. Science & diplomacy. 2019 May;8(1).

Bradshaw P, Hall J, Hill T, Mabelis J, Philo D. (2012). Growing up in Scotland: early experiences of primary school. Available from: https://www.gov.scot/publications/growing-up-scotland-early-experiences-primary-school/

Buck N, McFall S. Understanding Society: design overview. Longitudinal and Life Course Studies. 2011 Nov 21;3(1):5-17.

Cairns BJ, Liu B, Clennell S, Cooper R, Reeves GK, Beral V, Kuh D. Lifetime body size and reproductive factors: comparisons of data recorded prospectively with self reports in middle age. BMC medical research methodology. 2011 Dec;11(1):1-3.

Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care. data ran into trouble. Journal of medical ethics. 2015 May 1;41(5):404-9.

Caspi A, Houts RM, Belsky DW, Harrington H, Hogan S, Ramrakha S, Poulton R, Moffitt TE. Childhood forecasting of a small segment of the population with large economic burden. Nature human behaviour. 2016 Dec 12;1(1):1-0.

Chambers R. Should the Census Have More Spine?. Harvard Data Science Review.;2(1).

Chen Y, Aldridge T, UK-COVID19 National Core Studies Consortium, Ferraro CF, Khaw F-M. COVID-19 outbreak rates and infection attack rates associated with the workplace: a descriptive epidemiological study. Preprint ahead of publication. MedRxiv: https://doi.org/10.1101/2021.05.06.21256757

Chowdry, H (2018) 'Estimating the prevalence of the 'Toxic Trio': Evidence from the Adult Psychiatric Morbidity Survey'. Available from: https://www.childrenscommissioner.gov.uk/wp-content/uploads/2018/07/Vulnerability-Technical-Report-2-Estimating-the-prevalence-of-the-toxic-trio.pdf

Cornish R, Tilling K, Boyd A, Macleod J, Van Staa T. Using linkage to electronic primary care records to evaluate recruitment and nonresponse bias in the Avon Longitudinal Study of Parents and Children. Epidemiology (Cambridge, Mass.). 2015 Jul;26(4):e41.

Cornish RP, Tilling K, Boyd A, Davies A, Macleod J. Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years. International journal of epidemiology. 2015 Jun 1;44(3):937-45.

Cornish R. (2019). Using linked health and administrative data to reduce bias due to missing data and measurement error in observational research (Doctoral dissertation, University of Bristol).

Cornish RP, Macleod J, Boyd A, Tilling K. Factors associated with participation over time in the Avon Longitudinal Study of Parents and Children: a study using linked education and primary care data. International journal of epidemiology. 2021 Feb;50(1):293-302.

Cox F, Marshall A. (2017). Using the Census Longitudinal Studies for research on health and health inequalities.

Coulter, A., A. Busby, R. Giles, L. Wiginton, E. King and L. Williams (2020a) 'What does the public think about the use of administrative data to make longitudinal research more inclusive?' Public dialogue report**,** June 2020.

Cruise S, Kee F, editors. Early Key Findings from a Study of Older People in Northern Ireland: The NICOLA Study. Northern Ireland Cohort for the Longitudinal Study of Ageing, Centre for Public Health, Queen's University Belfast; 2017 Nov.

Clemens S, Gilby N. Life Study: Birth Component: Pilot: Face-to-face fieldwork. In Dezateux C and Elias P (Eds). (2016). Life Study: Birth Component: Pilot: Face-to-face fieldwork. DOI: 10.14324/000.wp.1485698

Dibben, C; Shuttleworth, I; Duke-Williams, O; Shelton, N; (2017) 9. Longitudinal studies in the United Kingdom. In: Stillwell J, editor. The Routledge handbook of census resources, methods and applications: Unlocking the UK 2011 census. Routledge; 2017 Aug 24.

Dingwall R, Iphofen R, Lewis J, Oates J, Emmerich N. Towards Common Principles for Social Science Research Ethics: A Discussion Document for the Academy of Social Sciences', Finding Common Ground: Consensus in Research Ethics Across the Social Sciences (Advances in Research Ethics and Integrity, Volume 1).

Douglas E, Rutherford A, Bell D. Pilot study protocol to inform a future longitudinal study of ageing using linked administrative data: Healthy AGeing in Scotland (HAGIS). BMJ open. 2018 Jan 1;8(1):e018802.

Elias, P. (2021) 'Promoting public engagement with longitudinal research', Warwick Institute for Employment Research, University of Warwick.

Elger B. Ethical issues of human genetic databases: a challenge to classical health research ethics?. Routledge; 2016 May 13.

Elliot M, Mackey E, O'Hara K, Tudor C. (2016) 'The Anonymisation Decision-making Framework'. UK Anonymisation Network, University of Manchester.

Elliot M, O'hara K, Raab C, O'Keefe CM, Mackey E, Dibben C, Gowans H, Purdam K, McCullagh K. Functional anonymisation: Personal data and the data environment. Computer Law & Security Review. 2018 Apr 1;34(2):204-21.

Evans H. Using data in the NHS: the implications of the opt-out and the GDPR. King's Fund website. www. kingsfund. org. uk/publications/using-data-nhs-gdpr. 2018.

Felitti VJ, Anda RF, Nordenberg D, Williamson DF, Spitz AM, Edwards V, Koss MP, Marks JS. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: The Adverse Childhood Experiences (ACE) Study. American journal of preventive medicine. 2019 Jun 1;56(6):774-86.

Ford DV, Jones KH, Verplancke JP, Lyons RA, John G, Brown G, Brooks CJ, Thompson S, Bodger O, Couch T, Leake K. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC health services research. 2009 Dec 1;9(1):157.

Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R, Allen NE. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. American journal of epidemiology. 2017 Nov 1;186(9):1026-34.

Fuchs D, Burnside L, DeRiviere L, Brownell M, Marchenski S, Mudry A, Dahl M, Chudley A, Longstaffe S, Hanlon-Dearman A. Report Title: The Economic Impact of Children in Care with FASD and Parental Alcohol Issues Phase II: Costs and Service Utilization of Health Care, Special Education, and Child Care.

Gambaro L, Joshi H. Moving home in the early years: what happens to children in the UK?. Longitudinal and Life Course Studies. 2016 Jul 18;7(3):265-87.

Galea S, Tracy M. Participation rates in epidemiologic studies. Annals of epidemiology. 2007 Sep 1;17(9):643-53.

Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, Smith P, Dibben C, Goldstein H. GUILD: GUidance for information about linking data sets. Journal of Public Health. 2018 Mar 1;40(1):191-8.

Gomes D. (2020) Novel Health Record Linkages. London: UCL Centre for Longitudinal Studies. Available from: https://esrc.ukri.org/files/funding/funding-opportunities/danielle-gomes-novel-health-record-linkages/

Gray L, Gorman E, White IR, Katikireddi SV, McCartney G, Rutherford L, Leyland AH. Correcting for non-participation bias in health surveys using record-linkage, synthetic observations and pattern mixture modelling. Statistical methods in medical research. 2020 Apr;29(4):1212-26.

Green F, Anders J, Henderson M, Henseke G. Who Chooses Private Schooling in Britain and Why? 2017. Available from: https://www.llakes.ac.uk/sites/default/files/Green%2C%20Anders%2C%20Henderson%20%26%20Henseke.pdf

Green J, Reeves GK, Floud S, Barnes I, Cairns BJ, Gathani T, Pirie K, Sweetland S, Yang TO, Beral V. Cohort profile: the million women study. International journal of epidemiology. 2019 Feb 1;48(1):28-9e.

Goldstein, H., Lynn, P., Muniz-Terrera, G. & Hardy, R., O'Muircheartaigh, C., Skinner, C. & Lehtonen, R., "Population sampling in longitudinal surveys (comment and debate)". Longitudinal and Life Course Studies, 6, 447 – 475. http://dx.doi.org/10.14301/llcs.v6i4.345

Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking data for mothers and babies in de-identified electronic health data. PloS one. 2016;11(10).

Houtepen LC, Heron J, Suderman MJ, Tilling K, Howe LD. Adverse childhood experiences in the children of the Avon Longitudinal Study of Parents and Children (ALSPAC). Wellcome open research. 2018;3.

Jay MA, Mc Grath-Lone L, Gilbert R. Data resource: the National Pupil Database (NPD). International Journal of Population Data Science. 2019 Mar 20;4(1).

Junghans, C., Feder, G., Hemingway, H., Timmis, A. and Jones, M. (2005) Recruiting patients to medical research: double blind randomised trial of opt-in versus opt-out strategies. British Medical Journal, 331,940.

Katz A, Enns J, Smith M, Burchill C, Turner K, Towns D. Population Data Centre Profile: The Manitoba Centre for Health Policy. International Journal of Population Data Science. 2019;4(2).

Keiding N, Louis TA. Perils and potentials of self-selected entry to epidemiological studies and surveys. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2016 Feb;179(2):319-76.

Kispeter, E. (2019) 'Public support for accessing and linking data about people from various sources: Literature review', Warwick Institute for Employment Research, University of Warwick.

Knoppers BM. Framework for responsible sharing of genomic and health-related data. The HUGO journal. 2014 Dec 1;8(1):3.

Lancaster C, Koychev I, Blane J, Chinner A, Wolters L, Hinds C. The Mezurio smartphone application: Evaluating the feasibility of frequent digital cognitive assessment in the PREVENT dementia study. medRxiv. 2019 Jan 1:19005124

Lavrakas PJ. Encyclopedia of survey research methods. Sage Publications; 2008 Sep 12.

Lawler M, Morris AD, Sullivan R, Birney E, Middleton A, Makaroff L, Knoppers BM, Horgan D, Eggermont A. A roadmap for restoring trust in Big Data. The Lancet. Oncology. 2018 Aug;19(8):1014.

Lemos, G. and Frakenburg, S. (2015). Trends and Friends: Access, use, and benefits of digital technology for homeless people and ex-homeless people. [online] London: Lemos and Crane. Available at: https://lankellychase. org.uk/wp-content/uploads/2015/01/Trendsand-Friends-2015.pdf [Accessed 4 Apr. 2019].

Leon DA, Lawlor DA, Clark H, Macintyre S. Cohort profile: the Aberdeen children of the 1950s study. International journal of epidemiology. 2006 Jun 1;35(3):549-52.

Libby G, Smith A, McEwan NF, Chien PF, Greene SA, Forsyth JS, Crombie IK, Macdonald TM, Morris AD. The Walker Project: a longitudinal study of 48 000 children born 1952–1966 (aged 36–50 years in 2002) and their families. Paediatric and perinatal epidemiology. 2004 Jul;18(4):302-12.

Lynn P. Methods for longitudinal surveys. Chichester: Wiley; 2009 Jan 26.

Lynn, P. (2015) “The Need for Representative Samples” in Goldstein, H., Lynn, P., Muniz-Terrera, G. & Hardy, R., O’Muircheartaigh, C., Skinner, C. & Lehtonen, R.. Population sampling in longitudinal surveys debate. Longitudinal and Life Course Studies, 6, 447 – 475. http://dx.doi.org/10.14301/llcs.v6i4.345,

Lynn P, Borkowska M. Some indicators of sample representativeness and attrition bias for BHPS and understanding society. Colchester, UK: Institute for Social and Economic Research, University of Essex. 2018 Jan 29.

Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, Handal M, Haugen M, Høiseth G, Knudsen GP, Paltiel L. Cohort profile update: the Norwegian mother and child cohort study (MoBa). International journal of epidemiology. 2016 Apr 1;45(2):382-8.

Mars B, Cornish R, Heron J, Boyd A, Crane C, Hawton K, Lewis G, Tilling K, Macleod J, Gunnell D. Using data linkage to investigate inconsistent reporting of self-harm and questionnaire non-response. Archives of Suicide Research. 2016 Apr 2;20(2):113-41.

McGrail, K; Jones, K; Akbari, A; Bennett, T; Boyd, A; Carinci, F; Cui, X; Denaxas, S; Dougall, N; Ford, D; Kirby, RS; Kum, H-C; Moorin, R; Moran, R; O'Keefe, C; Preen, D; Quan, H; Sanmartin, C; Schull, M; Smith, M; Williams, C; Williamson, T; Wyper, G; Kotelchuck, M. A Position Statement on Population Data Science: The science of data about people. International Journal of Population Data Science. 2018 Sep 10;3(4).

McKinstry B, Sullivan FM, Vasishta S, Armstrong R, Hanley J, Haughney J, Philip S, Smith BH, Wood A, Palmer CN. Cohort profile: the Scottish Research register SHARE. A register of people interested in research participation linked to NHS data sets. BMJ open. 2017 Feb 1;7(2).

Menni, C., Valdes, A.M., Freidin, M.B. et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med 26, 1037–1040 (2020).

Millett C, Zelenyanszki C, Binysh K, Lancaster J, Majeed A. Population mobility: characteristics of people registering with general practices. Public Health. 2005 Jul 1;119(7):632-8.

Milne BJ, Atkinson J, Blakely T, Day H, Douwes J, Gibb S, Nicolson M, Shackleton N, Sporle A, Teng A. Data Resource Profile: The New Zealand Integrated Data Infrastructure (IDI). International journal of epidemiology. 2019 Jun 1.

Mostafa T, Ploubidis G. Millennium cohort study. Sixth Survey 2015-2016: Technical report on response (Age 14). London: Centre for Longitudinal Studies; 2017 Feb.

Northstone K, Lewcock M, Groom A, Boyd A, Macleod J, Timpson N, Wells N. The Avon Longitudinal Study of Parents and Children (ALSPAC): an update on the enrolled sample of index children in 2019. Wellcome open research. 2019;4.

Olsen J, Melbye M, Olsen SF, Sørensen TI, Aaby P, Nybo Andersen AM, Taxbøl D, Hansen KD, Juhl M, Schow TB, Sørensen HT. The Danish National Birth Cohort-its background, structure and aim. Scandinavian journal of public health. 2001 Oct;29(4):300-7.

Openshaw S. Ecological fallacies and the analysis of areal census data. Environment and planning A. 1984 Jan;16(1):17-31.

Martin J, Tilling K, Hubbard L, Stergiakouli E, Thapar A, Davey Smith G, O'Donovan MC, Zammit S. Association of genetic risk for schizophrenia with nonparticipation over time in a population-based cohort study. American Journal of Epidemiology. 2016 Jun 15;183(12):1149-58.

McMahon R, LaHache T, Whiteduck T. “Digital Data Management as Indigenous Resurgence in Kahnawà:ke,” International Indigenous Policy Journal 6, no. 3 (2015): 6, https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1226&context=iipj.

Mindell JS, Tipping S, Pickering K, Hope S, Roth MA, Erens B. The effect of survey method on survey participation: Analysis of data from the Health Survey for England 2006 and the Boost Survey for London. BMC medical research methodology. 2010 Dec 1;10(1):83.

Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. International journal of epidemiology. 2018 Feb 1;47(1):226-35.

Neale B. Qualitative longitudinal research: An introduction to the Timescapes methods guides series. Timescapes Methods Guides Series, Guide. 2012(1).

Pell J, Valentine J, Inskip H. One in 30 people in the UK take part in cohort studies. The Lancet. 2014 Mar 22;383(9922):1015-6.

Plewis I, Calderwood L, Hawkes D, Hughes G, Joshi H. Millennium Cohort Study: technical report on sampling. London: Centre for Longitudinal Studies. 2007 Jul.

Pritchard, V. ‘Integration of Data Science in the Primary and Secondary Curriculum’. Royal Society, London, UK. 2018.

Pyper, D. The Public Sector Equality Duty and Equality Impact Assessments. Briefing Paper. Number 06591, 8 March 2018. House of Commons Library. London, UK.

Randall S, Brown AP, Ferrante AM, Boyd JH. Privacy preserving linkage using multiple dynamic match keys. International Journal of Population Data Science. 2019 May 23;4(1).

Schnell R, Borgs C. Proof of Concept for a Privacy Preserving National Mortality Register. International Journal of Population Data Science. 2018 Aug 29;3(4).

Schoeni RF, Stafford F, McGonagle KA, Andreski P. Response rates in national panel surveys. The Annals of the American Academy of Political and Social Science. 2013 Jan;645(1):60-87.

Scottish Centre for Administrative Data Research. (2018). Researcher Handbook: Data Linkage and administrative data research in Scotland. Available from: https://www.scadr.ac.uk/sites/default/files/Researcher%20access%20handbook%20Nov%202019%20-%20SCADR.pdf

Shlomo N, Skinner C, Kim MS. Theoretical Sampling Design Options for a New Birth Cohort: An Accelerated Longitudinal Design Perspective. Research Report, August 2019. Manchester: University of Manchester.

Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ, Dominiczak AF, Fitzpatrick B, Ford I, Jackson C, Haddow G, Kerr S, Lindsay R, McGilchrist M, Morton M, Murray G, Palmer CNA, Pell JP, Ralston SH, St Clair D, Sullivan F, Watt G, Wolf R, Wright A, Porteous D, Morris AD. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. BMC medical genetics. 2006 Dec;7(1):74.

Speight S, Maisey R, Chanfreau J, Haywood S, Lord C, Hussey D. Study of early education and development: baseline survey of families. Research report, July 2015. London: Department for Education.

Steptoe A, Breeze E, Banks J, Nazroo J. Cohort profile: the English longitudinal study of ageing. International journal of epidemiology. 2013 Dec 1;42(6):1640-8.

Sullivan, A., Joshi, H. and Williams, J. (2020) New birth cohort study: theoretical sampling design options, CLS Working Paper 2020/4. London: UCL Centre for Longitudinal Studies.

Taylor AE, Jones HJ, Sallis H, Euesden J, Stergiakouli E, Davies NM, Zammit S, Lawlor DA, Munafò MR, Davey Smith G, Tilling K. Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. International journal of epidemiology. 2018 Aug 1;47(4):1207-16.

Teague S, Youssef GJ, Macdonald JA, Sciberras E, Shatte A, Fuller-Tyszkiewicz M, Greenwood C, McIntosh J, Olsson CA, Hutchinson D. Retention strategies in longitudinal cohort studies: a systematic review and meta-analysis. BMC medical research methodology. 2018 Dec;18(1):1-22.

Tillin T, Forouhi NG, McKeigue PM, Chaturvedi N. Southall And Brent REvisited: Cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins. International journal of epidemiology. 2012 Feb 1;41(1):33-42

Tunstall H, Pickett K, Johnsen S. Residential mobility in the UK during pregnancy and infancy: are pregnant women, new mothers and infants 'unhealthy migrants'?. Social science & medicine. 2010 Aug 1;71(4):786-98

Vallance P. Providing evidence to the Houses of Commons Science and Technology Committee on the UK Science, Research and Technology Capability and Influence in Global Disease Outbreaks. July 2020. Available from: https://committees.parliament.uk/oralevidence/701/html/

Verschuren WM, Blokstra A, Picavet HS, Smit HA. Cohort profile: the Doetinchem cohort study. International journal of epidemiology. 2008 Dec 1;37(6):1236-41

Watson, N., & Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In P. Lynn (Ed.), Methodology of Longitudinal Surveys (pp. 157-182). Chichester: Wiley

Waugh A, Clarke A, Knowles J, Rowley D. Health and Homelessness in Scotland. Scottish Government, Edinburgh. 2018.

Wellcome's Longitudinal Population Studies Working Group. (2017) Longitudinal Population Studies Strategy. Wellcome Trust. London, UK. Available from: https://wellcome.ac.uk/sites/default/files/longitudinal-population-studies-strategy_0.pdf

Wolke D, Waylen A, Samara M, Steer C, Goodman R, Ford T, Lamberts K. Selective drop-out in longitudinal studies and non-biased prediction of behaviour disorders. The British Journal of Psychiatry. 2009 Sep;195(3):249-56.

Wood A, Denholm R, Hollings S, Cooper J, Ip S, Walker V, Denaxas S, Akbari A, Banerjee A, Whiteley W, Lai A. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. bmj. 2021 Apr 7;373.

Abbreviations

ACEs Adverse Childhood Experiences
ADRUK Administrative Data Research UK
ADS Administrative Data Spine
Census LSs Census Longitudinal Studies
CIS The Department for Work and Pensions Customer Information System
CLS Centre for Longitudinal Studies
DEA The Digital Economy Act (2017)
DNBC Danish National Birth Cohort Study
DPUK Dementias Platform UK
DWP Department for Work and Pensions
eDRIS Public Health Scotland's electronic Data Research and Innovation Service
EIA Equality Impact Assessment
EHRC Equality and Human Rights Commission
ESRC Economic and Social Research Council
FASD Fetal Alcohol Spectrum Disorder\
HDRUK Health Data Research UK
HESA Higher Education Statistics Agency
ID Identification/Identity
IDI Integrated Data Infrastructure
LPS Longitudinal Population Studies
MCS Millennium Cohort Study
METADAC Managing ethico-social technical and administrative issues in data access
MRC Medical Research Council
NHS Act The National Health Service Act 2006
NILS Northern Ireland Longitudinal Studies
NISRA Northern Ireland Statistics and Research Agency
NPD National Pupil Database
NRS National Records Scotland
ODISSEI Open Data Infrastructure for Social Science and Economic Innovations
ONS Office for National Statistics
PBPP Public Benefit and Privacy Panel for Health and Social Care
PEARL Project to Enhance ALSPAC through Record Linkage
PDS Patient Demographic Service
PRUK Population Research UK
PSED Public Sector Equality Duty
PSHE Personal, Social and Health Education
RCT Randomised Control Trial
RCUK Research Councils UK
SABRE Southall and Brent Revisited
SAGE Scientific Advisory Group for Emergencies
SAIL Secure Anonymised Infrastructure for Linkage
SEED Study of Early Education and Developmen
SLS Scotland Longitudinal Studies
SRSA The Statistics & Registration Service Act 2007
Stats NZ Statistics New Zealand
TEDS Twins Early Development Study
TPP Trusted Third Party
TRE Trusted Research Environment
UKDS UK Data Service
UKHLS UK Household Longitudinal Study
UK LLC UK Longitudinal Linkage Collaboration
UKRI UK Research and Innovation Council
UKSeRP UK Secure e-Research Platform
UPRN Unique Property Reference Number
WMA World Medical Association

Contributors

With thanks to those who were consulted and contributed to this Scoping Study.

Administrative Data Research - Scotland
Carol Morris; Chris Dibben

Administrative Data Research - NI
Dermot O'Reilly

Administrative Data Research UK
Emma Gordon; Paul Jackson

Avon Longitudinal Study of Parents and Children
John Macleod; Nic Timpson

Born in Bradford / Born in Bradford Better Start
John Wright; Rosie McEachan; Sally Bridges

Centre for Environmental Health and Sustainability
Anna Hansell; John Gulliver

Centre for Longitudinal Studies
George Ploubidis; Lisa Calderwood

Dunedin Cohort Study / eRISK Cohort
Terrie Moffit

Education Policy Institute
Leon Feinstein

The Electoral Commission
Mark Williams

English Longitudinal Study of Ageing
James Banks

Erasmus University
Tom Emery

ESRC Timescapes Initiative
Bren Neale

HeLEX
Miranda Mourby

Generation Scotland
David Porteous

Health Data Research UK
Andrew Morris; David Seymour; Rhos Walker

Health and Social Care NI
Martin Maycock

Healthy Aging in Scotland
Elaine Douglas

La Trobe University
James Boyd

Manitoba Centre for Health Policy
Mark Smith

Medical Research Council
Alex Bailey; Joe McNamara; Sarah Dickson

Northern Ireland Cohort of the longitudinal study of Ageing
Frances Burns; Frank Kee

NHS Digital
Garry Coleman

Office for National Statistics
Hannah Finselbach; Myer Glickman; Pete Stokes; Rose Elliot; Sarah Henry

UCL Institute of Child Health
Harvey Goldstein; Katie Harron; Ruth Gilbert

MRC Integrative Epidemiology Unit
Kate Tilling

Public Health Wales
Alisha Davies; Jiao Song

SAIL Databank/ Administrative Data Research - Wales
Chris Orton; David Ford

UK Data Archive
Matthew Woollard

UK Household Longitudinal Study
John Burton; Michaela Benzeval; Peter Lynn

The Wellcome Trust
Bruna Galobardes; Erica Pufall; Mary De Silva

Appendix 1: Scoping Study Methodology

Information gathering of population databases for research across the UK

A1.1 The scoping project has issued information requests to the existing ADRUK Research Centres in Scotland, Wales and Northern Ireland, asking about data sources, infrastructures and key experts across the devolved authorities. Equivalent approaches have been made to the ONS and NHS Digital as organisations owning large population databases (the ONS census register and the NHS patient register). Completed proformas have been received from each source.

Expert interviews

A1.2 The scoping study sought views and evidence from representatives of:

- UK and international longitudinal studies (those supported by both the ESRC and biomedical funders);
- Academic data science networks and infrastructure providers, governmental data owners and infrastructure providers (at a senior policy level and expert insights from those familiar with coverage, quality and data management);
- Experts in statistical methodologies for dealing with missing data and bias and experts in data sharing law and governance.

Stakeholders were selected from across all four UK nations, with some international experts where relevant. Initial selection of interview candidates was role based (i.e., LPS principal investigators, policy directors and data experts in health and government departments) with subsequent interview candidates identified through recommendations using a snowball sampling strategy (for a list of those interviewed see **Contributors** statement above).

Desk-based research

A1.3 The desk-based research sought to identify existing evidence describing challenges and successes within UK LPS sampling, recruitment and ongoing participant retention activities. The scoping study did not have the resources to conduct a systematic literature review, so adopted an overview literature review strategy to search on the names of the UK LPS and the terms ‘selection’, ‘attrition’, ‘representation’ and, ‘bias’. This was designed to complement key literature identified through the expert interviews. Much of the evidence of interest is methodological and may be reported in conference procedures and may be underrepresented in peer-reviewed journals. To account for this, the conference procedures (abstract books) of conferences attended by LPS and data scientists were systematically searched using the keywords: ‘sampling frame’, ‘selection’, ‘attrition’, ‘representation’, ‘bias’ and ‘follow-up’ / ‘follow up’. The proceedings of the ‘Longitudinal and Lifecourse studies’, ‘European Social Research Association’, ‘International Population Data Linkage Network’, the ‘Farr Institute’ and ‘SHIP’ conferences were searched.

A1.4 The desk-based research is intended to identify existing population registers (at a UK level and devolved administration level) and in particular academic and government initiatives which are bringing together routine records from different sources.

Commissioned evidence pieces

A1.5 The HeLEX group at University of Oxford have been commissioned to provide opinions on the potential legal basis for combining population identifiers from different sources and using these for population research, and on the relevance of Equality legislation to the process of commissioning and supporting LPS and of the process of designing the sampling and recruitment and follow-up strategies of LPS.

A1.6 Two research projects have been commissioned which seek to investigate the legal, governance and technical pathways for combining diverse data across city/regions. These exercises will identify the key stakeholders and decision makers involved in population data science and resource building at a local level. As such they will identify the potential for developing local/regional population databases. The exemplar projects are set within the Manchester devolved health authority and within the Bristol, North Somerset and South Gloucestershire (BNSSG) NHS catchment. The Manchester exemplar is focused on data integration across health and social providers, whereas the BNSSG exemplar is focused on linking an LPS population into a whole population database within the same geographical area.

Appendix 2: Example LPS sampling approaches

Avon Longitudinal Study of Parents and Children (ALSPAC)

All pregnant women who were due to deliver between 01.April.1991 and 31.December.1992 while living in and around the City of Bristol were deemed eligible for inclusion in the study. Recruitment was conducted through publicity campaigns, by midwives and by study staff at antenatal scan clinics and on the delivery wards. Recruiting a prospective birth cohort during pregnancy has the particular challenge of not having a defined sampling frame: as not all women book for maternity care, that many book at different stages of the pregnancy, and during ALSPAC's recruitment, maternity care records were paper-based and distributed across midwifery teams. The eligible study sample has been retrospectively defined, based on an estimate of the true eligible sample which was achieved through linking ALSPAC recruitment records to maternity, birth and child health services records (Boyd *et al*, 2013). This has enabled the assessment of recruitment coverage (Boyd *et al*, 2013) and recruitment bias (Cornish *et al*, 2015). ALSPAC secured the support from a Health Research Authority (HRA) Research Ethics Committee and the HRA Confidentiality Advisory Group to access the records of eligible participants (eligible under the original study criteria), who had not been contacted during the original recruitment campaign in order to contact them via a postal campaign and to invite them to enrol (Northstone *et al*, 2019). ALSPAC were only permitted to use the contact details for this recruitment campaign and they were not integrated into the main study database, unless a participant consented to enrol.

English Longitudinal Study of Aging (ELSA) & the Health Survey for England (HSE)

The eligible sample was drawn from participants in the 1998, 1999 and 2001 sweeps of the cross-sectional Health Survey for England (HSE) and their partners. In addition, the index participant needed to have been born before 01/05/1952 and live in a private household (Steptoe *et al*, 2013). Additional samples of individuals aged 50–75 years were added at waves 3, 4 and 6 from the 2001-2004, 2006 and 2009-2011 HSE sweeps. In turn, the HSE (2018 sweep) is intended to be a representative sample of private households in England. The sampling used a multi-stage stratified probability design. This comprised a random selection of geographical units (based on postcode sectors) from the Postcode Address File; from which, a random sample of postal addresses was drawn. Most adults and some children in the household are then invited to be surveyed. An initial letter was sent by post accompanied by a £10 voucher. This was followed-up by a fieldworker visit/interview. Boost samples have been used in different waves (Mindell et al, 2010).

Generation Scotland: the Scottish Family Health Study

Generation Scotland is a family study with a focus on genetic epidemiology. During phase 1 (2006-10) of recruitment, all individuals registered with a participating GP practice (members of the Scottish Practices and Professionals Involved in Research (SPPIRe) network) in the Glasgow and Tayside area and aged 35-65 were considered eligible. Those eligible were screened by their GP, who removed those who lacked capacity to consent and those where it was considered inappropriate to recruit (e.g., those with serious or terminal illness). The sample was increased through including traceable members of the Walker Birth Cohort Study: a records-based cohort of 48000 individuals born in Dundee between 1952 and 1966 (Libby *et al*, 2004). Those eligible were invited through postal invitation and were asked to enrol on the proviso that they recruited at least one other family member aged 18+. In phase

II (2011) of recruitment, the catchment area was expanded to include Ayrshire, Arran and Northeast Scotland and the age range adjusted to 18-65 (Smith *et al*, 2013).

Healthy AGeing in Scotland (HAGIS)

HAGIS is a longitudinal study of aging in Scotland, as such is a 'sister' study to ELSA (England), NICOLA (NI), TILDA (Republic of Ireland), Health and Retirement Study (USA). A two-stage cluster-sampling approach was used to randomly sample residential addresses (using the Postcode Address File) stratified by geographical area, urban/rural status and Scottish Indices of Multiple Deprivation (Douglas *et al*, 2018). The National Records of Scotland screened the sampled addresses to identify those with a resident aged 50+ using information from the NHS Scotland Central Register. An invitation letter was sent to all sampled addresses with study information and an opt-out form. After 7-10 days, this was followed-up by a fieldworker visit (from an outsourced agency) unless an opt-out was received. Permission for this use of data was granted by the Public Benefits and Privacy Panel (PBPP).

Millennium Cohort Study (MCS)

The recruitment to MCS aimed to provide data about a representative sample of children in each of the four countries of the UK and to include 'usable' (i.e., sufficiently powered) data for sub-groups of children living in advantaged and disadvantaged circumstances; of ethnic minorities and those living in each of Scotland, Wales and Northern Ireland (Plewis *et al*, 2007). A probability (or random) method of sample selection combined with stratification and clustering was adopted. The strata were established using aggregate population data at electoral ward level: with every ward allocated to an 'advantaged', 'disadvantaged' and, within England only, a 'ethnic minority' strata. Eligible wards were selected systematically against a set of ideal recruitment targets within each stratum, country, and then for England and Scotland, region. All births in selected wards were considered eligible if the child was born between 01.September.2000 - 31.August.2001 (England and Wales) and 24.November.2000 - 11.January.2002 (Scotland and Northern Ireland), and that they were alive, living in the UK and eligible to receive Child Benefit at age nine months. A list of all the selected children was generated using information within the register used to administer the Child Benefit, which was at the time a universal benefit typically paid to the child's mother by the UK Department of Social Security (DSS). The DSS wrote to all selected mothers providing information about the study and inviting them to take part. An opt-out was offered at this point. Where there was no opt-out, the contact details were passed to the study team who in turn contracted a fieldwork agency to contact the families in a face-to-face visit to recruit the child and collect data. In theory the sample included vulnerable children living outside of traditional 'households' (e.g., women's refuges, hostels, hospitals, prisons). However, the sample issued by the DSS was filtered to exclude individuals who were deemed to have sensitive circumstances (including all children in State Care, where there had been a death in the family in the last five years, or where the family were in correspondence with the DSS).

Northern Ireland Cohort for the Longitudinal Study of Aging (NICOLA)

The NICOLA aging cohort is designed to be complementary with other 'sister' studies and for this reason adopted a similar sampling strategy (Cruise and Kee *Eds*, 2017). Eligibility is defined as all individuals aged 50+ and living in private residential accommodation in Northern Ireland. Attribute information (age) and residential address details were extracted

from GP registrations; stratified by geographical area and then systematically (fixed interval) sampled. The initial sampling included the primary frame and a reserve, with an additional top-up sample selected in a second exercise. Households were sent an unnamed letter which was then followed up with a phone call from the IPSOS MORI field interviewer to arrange a home visit. Individuals could opt-out at this stage over the phone. The sample selection excluded individuals living outside of private residences and those lacking capacity to consent.

Study of Early Education and Development (SEED)

SEED is a Department for Education commissioned cohort study designed to investigate the impact of early years education interventions (funded pre-school teaching). The study recruited young children (age ~2 years) and families across a spectrum of socio-economic status. The sampling frame (Speight *et al*, 2015) was compiled from DWP child benefit information (notably, after changes to child benefits which meant it was no longer a universal benefit and thus no longer applied to higher-income earning families). Sampling was geographically clustered for efficiency and to encourage overlap of use of the same early years and childcare settings. Geographies were selected by postcode districts and sub-geographies by postcode sector. Families were then, at an individual level, allocated into one of three SES groupings by the DWP: 20% most disadvantaged; 20-40% moderately disadvantaged; not disadvantaged >40%. SES status was determined using a range of benefit provisions which in turn was based on household earnings data. Families were sent a DWP branded letter offering an opt-out to DWP providing contact information to the study: excluding those who opted out, contact details were provided to a fieldwork agency who then sent an advance letter which was followed-up by a fieldwork visit. Of those issued to the sample (n=9188), 98% were deemed eligible and 86% were contacted. 22% refused and there was an overall response rate of 63%. Response was lowest amongst the most disadvantaged families.

Southall and Brent Revisited (SABRE)

SABRE is a 20-year longitudinal follow-up of two cross-sectional samples designed to understand differential rates of diabetes, coronary heart disease and stroke amongst different ethnic migrant population groups: the Southall study (1988-1990) is an inter-ethnic cross-sectional study of middle-aged men and women in West London, UK; the Brent study comprised African Caribbean and European middle-aged men and women in North-West London, UK (Tillin *et al*, 2012). Both these areas have high levels of ethnic diversity and economic disadvantage. Southall participants were sampled from local factories and general practitioners' (GPs') registers. Brent was a stratified sample (on ethnicity and gender) selected from GPs' registers. The combined studies have 4972 enrolled participants from a sample of 7942 (63% recruitment rate). Participants were re-contacted for the 20-year follow-up using contact information (filtered for mortality status) from the NHS patient register.

UK Biobank

Eligibility for UK Biobank is defined as all individuals aged between 40-69, who were living within 25 miles of one of 22 study centres located across England, Wales and Scotland and who were registered with the NHS (Fry *et al*, 2017). Contact details and some attribute data were sourced from the NHS with support of the Health Research Authority. No filters (other than location) were applied to the sample selection. The HRA approvals permit limited,

anonymised, socio-demographic data to be retained on all eligible individuals (sex, month, and year of birth, Townsend deprivation index (an indicator of socioeconomic status), and geographic location).

UK Household Longitudinal Study (Understanding Society)

Understanding Society is a UK household longitudinal study with three primary components: a general population sample, an ethnic minority boost sample and the sample of participants from the British Household Panel Survey (BHPS) (Buck and McFall, 2011). In Great Britain, the general population sample is a proportionally stratified, clustered, equal probability sample of residential addresses. Using the Postcode Address File, postal sectors (a postal system area) were stratified by English regions and Scotland and Wales, occupational classification (using 2001 Census information) population density and minority ethnic density. The stratification variables (GOR, social class, population density and ethnic minority density) were chosen for their likely correlation with key survey measures. Within these strata, postal sectors were selected systematically, with probability proportional to size (number of addresses), and 18 properties were systematically chosen. In Northern Ireland, addresses were systematically selected from the Land and Property Services Agency domestic property register. The ethnic-minority boost sample was selected from geographic areas with at least 5% density of ethnic minority groups and sub-sampling to identify areas with higher density of ethnic minority households (Boreham *et al*, 2012). Fieldworkers screen households for eligibility. Understanding Society incorporates the BHPS sample households, which were recruited from a randomly selected sample of residential properties across Great Britain (excluding the Scottish Highlands and Islands) with subsequent boost samples in Scotland and Wales. There is an additional 'innovation panel' which was sampled from across the UK (excluding the Scottish Highlands and Islands).

Appendix 3: Population Databases defining Vulnerable sub-groups

Troubled Families programme (TF)

The TF Programme targets 'early' interventions for families with multiple problems (crime, anti-social behaviour, truancy, unemployment, mental health problems and domestic abuse) with an objective to drive change in local authorities to take a 'whole family approach' to providing support. Families are identified at a household level, using routine information from diverse routine records compiled by government service providers (e.g., social care, education, police and health records) augmented by third-sector records of organisations working specifically to help those who are vulnerable or marginalised. **Specific legislation was passed to enable the data sharing,[58] there is a discussion of data sources and their legal gateways elsewhere.[59]**

The Children's Commissioner: using population data to define vulnerability

A critical question for the Children's Commissioner is to identify how many children in England have needs that will create demand for statutory or acute care, looking in the round across government and local authority services (health, education, justice, social care, welfare and benefits). They have identified 70 indicators of vulnerability using all available public data. It is estimated that, in 2019, there are 1 million children needing help for mental health problems, 120,000 who are homeless and living in temporary accommodation, >50,000 children who are not receiving any education; and nearly 30,000 are in violent gangs. Some of these are receiving intensive state support, some an unclear level of support and a larger group are unknown to service providers. Over 1 million children have a long-term limiting illness (e.g., asthma, epilepsy, diabetes) and significant health inequalities exist within England.[60] An estimated 2.3 million children are living with risk because of a vulnerable family background where an adult in the household as one or more of the 'toxic trio' of domestic violence and abuse, substance misuse or suffers from mental health issues (Chowdry, 2018). Just over half of these are 'invisible' to service providers (i.e., they are not receiving services and their vulnerability is unrecorded). **These figures demonstrate the scale of 'vulnerability' and marginalisation within children in England, that many do not have official records indicating their vulnerability and that many are vulnerable due to the characteristics of those they live with.**

Adverse Childhood Experiences

ACEs can be defined using information from routine records or survey data (or both). Many ACE studies use retrospective data and where self-reported this can be susceptible to recall and selection bias. LPS data are well suited to the study of ACEs and their outcomes given that the data is collected across the lifecourse and ideally within samples where selection is controlled for. However, it can be seen that when pooling data to define ACE status, LPS can have high levels of missingness (Houtepen *et al*, 2019). There remains debate as to the

[58] The Social Security (Information-sharing in relation to Welfare Services etc.) Regulations 2012, SI 2012/1483.

[59] Financial Framework for the Troubled Families Programme. (2018). Department for Communities and Local Government. London, UK.

[60] Children's Commissioner's Briefing: Health Inequalities in Childhood. (2020). Office of the Children's Commissioner. London, UK.

relevant life course factors to identify which experiences should be considered when defining ACE status (for example, should low Socio-Economic Position be included in the framework given it is known to be strongly associated with overall ACE status?). Felitti (Felitti *et al*, 2019) demonstrated that adverse outcomes increased with increasing number of ACEs, and a cut of 4+ ACEs is now commonly used to define at risk individuals. **However, as with any threshold-based approach this assumes all ACEs are equal in their effects and does not take into account timing, duration or severity.**