Exploring Population Data for Inclusive Research: A scoping project

Initial findings

Andy Boyd, University of Bristol and Rebecca Perring, ESRC, August 2020

Background

The 2017 Longitudinal Studies Strategic Review (LSR) was commissioned to evaluate the continuing and future scientific needs, challenges and opportunities for longitudinal research resources in the UK. The LSR explored how ESRC could meet these needs and offered recommendations on strategic and innovative ways to enhance this portfolio of investments for the future.

The LSR identified the need to address key research challenges by improving the inclusivity of all sections of society in existing and new UK Longitudinal Population Studies (LPS). It identified the need for a new cohort and that existing problems in recruiting and retaining representative samples across LPS for any new and existing LPS needed to be addressed, recognising the risk that harder to reach, marginalised and vulnerable groups may be excluded from research. To address these issues the ESRC commissioned a scoping project, undertaken by Andy Boyd, University of Bristol, to explore potential options to address them. The initial findings are presented here, with the expectation that the full report will deliver later this year.

Scoping study remit

This scoping study gathered and distilled a range of evidence which explores how population data can be used to help ensure longitudinal research is inclusive. To achieve this aim, study samples need to be heterogenous, particularly of vulnerable and marginalised groups of people, who can be the 'hardest to reach'. These groups are often in need of resource intensive state support – while not always having those needs identified or met – and are therefore of particular interest to researchers and policy makers. The scoping study centred around two key objectives:

1. To identify the kinds of population data sets that are in operation in the UK, in order to understand if a new type of infrastructure/ methods might be needed for inclusive research and what it might look like;

2. To identify approaches for understanding the population coverage of UK longitudinal studies and to identify missing populations, particularly vulnerable and marginalised groups, in order to help ensure that any new cohort is inclusive and representative of the UK population, and that its coverage is understood; and to gather learning and practice that could be used by existing studies.

Key findings and recommendations

1. An 'administrative data spine' for research is not proportionate or acceptable

The LSR considered the potential value for recruitment and follow-up that could be delivered by an 'administrative data spine' (ADS). The study identified that a whole population ADS – a centralised, identifiable, whole population research register drawing data from across government departments - could deliver potentially meaningful scientific and efficiency benefits and is technically possible. However, at the time of scoping the study identified that an ADS model as a means to address the problems identified by the LS Review is likely to be neither feasible, proportionate nor acceptable. Some experts considered that the negative consequences for confidentiality and the cost of such an exercise would not be justified by the

benefits it would deliver and the public good it would serve. It was suggested that anticipated public distrust of, and hostility to, registry based ways of working in government/academia 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. **It was therefore recommended that the ESRC should not pursue this specific option at present.**

It is, however, important to note that this study was mostly undertaken before the COVID-19 pandemic. It is clear that new insights are emerging about the ways in which population data can be used to improve services and help target resources towards people most in need and these may have changed both the public's and data owners' appetite for making better use of population data in these ways. It is important that the benefits that an ADS might deliver, as highlighted by this study, form part of any future conversations about the potential for data to support public service delivery, statistics and research. The ESRC will contribute to these developing conversations.

At this time, in the absence of an ADS, the study identified that **the most suitable datasets for sampling and recruiting participants to LPS in the UK are currently the birth register and the national NHS Patient Registers**. These contain sufficient identifiers and contact details, have established precedents for being used in this way, have high population coverage (for birth cohorts), and exist in similar forms across the UK nations. However, it is notable that data access routes vary across the UK. These data could also lack some of the socio-economic indicators used in some study sampling designs and are likely to lack information needed to target recruitment to some specific vulnerable and marginalised sub-groups.

2. Sampling selection and recruitment should be undertaken at an individual level through a privacy preserving mechanism

The project found evidence that sampling and recruitment decisions are generally made at an area level rather than an individual level, introducing a risk that the achieved sample selection is not optimal and fully heterogeneous of the diversity reflected in the real world. Measures to assess inclusion and representation in the recruited sample are based on similarly aggregated data which may mask the full patterns of recruitment bias. The use of area level data appears a compromise driven by data availability and pragmatic and efficient approaches to fieldwork (i.e. the cost efficiencies of recruiting within defined geographical areas). There are also privacy issues and notions of acceptability for the use of individual data prior to recruitment, which are reflected in concerns raised by the public and privacy advocates, and in terms of the ethico-legal basis for releasing data. To address this issue, **it is recommended that a privacy preserving protocol is considered for sample selection and recruitment.** This is particularly the case if a study intended to over sample or stratify sampling using sensitive characteristics.

A privacy preserving protocol would enable LPS designers to establish a sampling frame using sufficient granular individual-level data to isolate as close as an approximation to the target population as is possible. It would also enable the release of sufficient information from the sampling frame to the recruiter to be able to make contact with the selected cases while minimising the disclosure of information to the recruiter. With investment, it would be technically feasible for recruiters to interact with underlying sensitive data in a manner in which an individual's information was not disclosed. This could help ensure that recruitment performance is balanced in line with the design requirements and that the fieldwork team is informed and can be responsive to results achieved in the field. The protocol should accommodate the variety of sampling designs used in UK LPS and would also be applicable to wider scenarios (e.g. recruitment into clinical trials) which may bring additional value to the development of such a resource.

3. There is a need to assess diversity and inclusion at both an individual LPS and community level

The study found that a disproportionate amount of those missing from LPS are vulnerable and marginalised. **It is recommended that diversity and inclusion are assessed at individual study level, and separately at the longitudinal community level** to consider how the sum total of LPS are representative and inclusive of the UK population.

To address this **LPS should develop and implement evidence based 'Inclusion plans'** and that these should be developed and refined with input from participants and members of the impacted sub-groups. **LPS funders should provide mechanisms to support studies' inclusion and engagement activities** and monitor delivery using relevant metrics. It was noted that successful engagement strategies of harder to reach participant groups are based on developing trust relationships and that this is a long-term endeavour. Short-term funding cycles or reprioritisation of resources represents a threat to the success of this, and unsuccessful attempts to build relationships could reinforce perceived feelings of marginalisation within sub-groups.

It is recommended that linkage informed strategies be considered to help understand inclusivity at community level. For example, the 'flagging and tracing' mechanism with NHS Digital – which is used by many UK LPS to match participants to their NHS register record to elicit health information and contact information- could be used to provide a means of defining the national population of individuals enrolled in a longitudinal study. This combined view, would provide a means of understanding participant characteristics and coverage of the estimated 2-3 million individuals participating in a UK LPS in comparison with the whole UK population. Further, a fuller range of available socio-economic status metrics could potentially be encompassed through the use of a heavily de-identified whole population data flag. Although there is currently no model for this, this could potentially be facilitated by ONS via the Digital Economy Act in conjunction with NHS Digital, or through equivalent data processing.

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

The study found that LPS have a social and ethical obligation to conduct high-quality research that – within the bounds of their study design - is inclusive of vulnerable groups and groups experiencing disadvantage(s) related to their protected characteristics. It is not the case that all LPS need be fully inclusive of all sub-groups, or to always have representative samples: sampling strategies should be driven by the scientific purpose of the given study. However, it is arguably the case that the LPS community should not systematically exclude (or fail to provide reasonable measures to facilitate inclusion) of vulnerable and marginalised groups. **It is recommended that there is an onus on research funders and those developing LPS strategic thinking to ensure inclusion, fairness and equality at the level of the longitudinal community.**

It is imperative when ensuring there is an appropriate legal and ethical basis for research, **that the public are empowered to help shape LPS** 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 license.

5. Long term follow-up of participants through linkage infrastructure

The study highlighted that barriers to accessing data and navigating changing governance frameworks have resulted in uneven development of the use of linked routine records in LPS. Yet where these have been successful, there is a growing portfolio of studies using linked records to quantify patterns of inclusion and

representation, and to address missing data and attrition bias. Barriers to entry may result in some studies – including those following vulnerable and marginalised sub-groups – not being able to implement linkages and therefore not maximising the value of their participants' contribution or the study's ability to fully inform policy development. **Community based approaches to record linkage should be considered** such as studies sharing insights, tools and materials to the co-development and management of new research infrastructure. **The jointly funded Population Research UK initiative could provide the framework for this.**

In terms of long-term linkage follow-up, a community consortium type model could potentially be acceptable to both studies, participants and other key stakeholders. This could deliver some of the benefits that an ADS might, such as using linked records to address missing survey data and to assess error and bias, to contribute new data not available directly from participants, and to calculate study weightings using linked records in addition to baseline study data.

To help ensure acceptability, this type of approach should retain study control mechanisms and be respectful of existing study operating frameworks and participant assurances. **The sharing of linked data should be transparent and occur using data de-identified to managers and research analysts within 'Trusted Research Environments' whose security and governance processes are accredited and subject to independent audit** (e.g. UK Data Archive, UK Secure eResearch Platform, ONS Secure Research Service). It is, however, important to recognise that this alternative to the 'population register' element of the ADS – compiling a list of the population through linking data together from across different data owners – may result in a resource which is not fully inclusive of the population.

It is recommended that the LPS community is encouraged and supported to consider and develop strategies for new ways of working in record linkage that enables successful creation of interdisciplinary linkages, and is structured and managed in a manner that is financially sustainable and responsive to changing expectations for governance and public acceptability. This needs to be led by individual studies to help ensure established participant-study trust relationships are not eroded.

Next steps

The use of population data in sampling, recruitment and retention will only partly help address the challenges identified. 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 are also effective and important to continue. **It is important that these functions receive continuing support and resources.**

This scoping project occurred during a period of time before the COVID-19 pandemic and before renewed thinking about inclusion and equality arising from the Black Lives Matter movement. UKRI's Black Lives Matter statement reflects on the research community's responsibilities to more in this respect. These developments, along with current and future 'shocks' – such as the climate crisis or future recessions, further pandemics or periods of abrupt technological or social changes - provide justification for the need to have a flow of information. This is important to enable ways of working which can facilitate responsive and inclusive interdisciplinary research which feeds into evidence-based policy development. **The LPS community should work with participants/public and experts, for example the Understanding Patient Data group, to explore how COVID-19 research can be used to exemplify the benefits of population data science and to determine if new ways of thinking about data use amongst study participants and the public in general can enable new ways of working in longitudinal studies.**

It is anticipated that the full report will be available later in the year.