Participant acceptability of ‘digital footprint’ data collection strategies: evidence from the index participants of the ALSPAC birth cohort study.

Andy Boyd[1,2], Kate Shiells[3,4], Nina Di Cara[5,6], Anya Skatova[3,4], Oliver S.P. Davis[4,6], Claire M.A. Haworth[3,4,6], Andy L Skinner[6,7], Richard Thomas[1], Alastair R Tanner[6], John Macleod[1,8], Nicholas J Timpson[1,6]

[1] Avon Longitudinal Study of Parents and Children, Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK; [2] CLOSER longitudinal study consortium, University College London, London, UK; [3] School of Psychological Science, University of Bristol, Bristol, UK; [4] Alan Turing Institute, London, UK; [5] Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK; [6] MRC Integrative Epidemiology Unit, Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK; [7] School of Experimental Psychology, University of Bristol, Bristol, UK; [8] NIHR Applied Research Collaboration West, Bristol, UK.

December, 2019

Acknowledgements

This evidence synthesis report was funded from UK government Strategic Priorities Fund and awarded by the Economic and Social Research Council (ESRC).

Avon Longitudinal Study of Parents and Children (ALSPAC)
We are extremely grateful to all the families who take part in the study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

The UK Medical Research Council and Wellcome (Ref:102215/2/13/2) and the University of Bristol provide core support for ALSPAC. The ALSPAC evidence in this report project draws on methodologies developed in the Project to Enhance ALSPAC through Record Linkage (PEARL) – a Wellcome Trust award (WT grant reference: WT086118/Z/08/Z).

For the research work that informed this synthesis: AB and RT were funded by the UK Natural Environment Research Council (R8/H12/83/NE/P01830/1) as part of the 'Enhancing Environmental data Resources In Cohort studies: ALSPAC exemplar' (ERICA) award; AS was funded by a University of Bristol Vice Chancellors Fellowship; KS, AS, OSPD, CMAH were funded by the Alan Turing Institute under the EPSRC grant EP/N510129/1; OSPD, CMAH and AB were funded by the ESRC (ES/R011583/1) as CLOSER Work Package 21; AB, JM, OSPD and CMAH were funded by an MRC Mental Health Data Pathfinder award (Ref: MC_PC_17210). ALS is supported by a HDR UK Fellowship; OSPD, CMAH and ALS work inca Unit that is supported by the UK Medical Research Council (MC_UU_00011/6 and MC_UU_00011/3) and University of Bristol. NJT is a Wellcome Trust Investigator (202802/Z/16/Z), is supported by the University of Bristol NIHR Biomedical Research Centre (BRC-1215–20011), and works within the CRUK Integrative Cancer Epidemiology Programme (C18281/A19169). The views expressed in this report are those of the authors and not necessarily any funding body. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the report.

ALSPAC
Oakfield House
Oakfield Grove
Bristol
BS8 2BN
United Kingdom
Tel: +44 (0) 0117 331 0010
Email: info@childrenofthe90s.ac.uk
Web: https://www.bristol.ac.uk/alspac/
Twitter: @CO90s

Citation of the Report:
Boyd A, Shiells K, Di Cara N, Skatova A, Davis OSP, Haworth CMA, Skinner AL, Thomas R, Tanner AR, Macleod J, Timpson NJ. (2019). Participant acceptability of 'digital footprint' data collection strategies: evidence from the ALSPAC birth cohort study. Bristol, UK: University of Bristol.

Executive Summary

- ALSPAC have identified that digital footprint records offer new research utility and are developing a strategy for incorporating these data into the study databank;
- It is imperative that this novel use of participant data is understood and acceptable in order to successfully establish linkage to these data whilst maintaining trust in the wider study;
- ALSPAC are testing participant views and expectations through qualitative research with participant groups to inform the design of our digital footprint strategy;
- The evidence collected to date suggests that participants are initially unfamiliar with the rationale for using these forms of data in the ALSPAC research programme, and are unsure as to what benefits this can bring;
- While some of these data are already in the public domain (e.g. twitter posts, air pollution measures), there is a perception that this is a substantial change in the studies data collection strategy;
- Some participants consider that some forms of digital footprint data are more sensitive than others: with information on detailed online and transactional behaviours considered particularly sensitive, as are precise location information, bank records and medical records;
- Participants consider that maintaining confidentiality, receiving clear information about any proposed data use and having mechanisms to retain control over the use of their data are important safeguards when considering requests to use these data;
- Once the value and benefits are made clear, and subject to safeguard controls being in place, then many – but not all – participants suggest they would be accepting of this use of their records;
- The use of digital footprint records will introduce ethical challenges, but the majority of these challenges align with existing ethical issues and can be accommodated within existing study frameworks.

Contents

1. Introduction

1.1 background

Almost everyone are now generating high resolution time- and place-stamped 'digital footprint' records during their routine daily actions and social interactions. The extent and manner of this record generation is vast and increasingly diverse: with records being generated as individuals shop, seek information, communicate and interact, consume entertainment and conduct routine activities in ways that are mediated by digital devices, involve digital services or take place within data gathering and connected environments.[1] Some of these data are created by the participant and some are generated automatically. Within this, an individual's digital footprint data can be generated actively (e.g. creating a social media post) or passively (e.g. through the tracking of online behaviour). The rates of population uptake of connected digital devices and digital services are increasingly high, for example in 2018/19:

- 92% of the UK population had used the internet;
- 84% of all UK adults were accessing the internet using mobile devices.
- 82% of all UK adults bought goods or services online (within the last 12 months);
- 63% of adults used the internet to look for health information;
- 23% of households owned at least one 'smart' device;

However, there are demographic variations in this – for example, 71% of UK adult women used social networking compared with 64% of men - and lower rates of uptake are found in some vulnerable groups (e.g. only 78% of disabled adults had used the internet by 2019).[2,3]

Digital footprint records provide an opportunity for high resolution and ecologically valid data on social interactions which in turn can deliver insights into health and social behaviours when linked to longitudinal population study (LPS) datasets. Furthermore, the systematically collected self-report data and linked objective records within LPS can provide opportunities to measure and assess the 'ground truth' about behaviours and outcomes which will bring insights into the validity of digital footprint sources in population research.

The value of LPS collecting data using linkage to novel records has been recognised across all major funders of longitudinal research in the UK.[4,5,6] This recognition encompasses well-established strategies to link to participants routinely generated health and social government department records and also the opportunities to link across a range of digital footprint records. Reflecting this, the Avon Longitudinal Study of Parents and Children (ALSPAC) birth cohort study is considering how these opportunities could be realised, and how linkage to digital footprint records would fit within the existing framework of study

[1] Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences. 2013 Apr 9;110(15):5802-5.
[2] Internet users, UK: 2019. 2019. ONS, London, UK.
[3] Internet access – households and individuals, Great Britain: 2019. 2019. ONS, London, UK.
[4] Pell JP, Valentine J, Inskip H. One in 30 people in the UK take part in cohort studies. Lancet (London, England). 2014 Mar 22;383(9922):1015.
[5] Longitudinal Population Studies Strategy. 2017. Wellcome Trust, London, UK.
[6] Longitudinal Studies Strategic Review. 2017. ESRC, Swindon, UK.

operations and governance. To inform this thinking, the study and collaborating researchers have been consulting participants in order to understand views relating to digital footprint data collection strategies and to ensure that learning gained from these consultations is fed into the design of linkage methodologies, governance frameworks and communication materials.

1.2 Scope of the report

This report provides a synthesis of participants views on the understanding and acceptability of digital footprint data collection within ALSPAC followed by evaluation of the ethical issues arising and the manner by which solutions to some of these challenges have been implemented into study practices. The evidence has been collected by the authors using qualitative interviews, focus groups or through consultation with participant advisors.

For the purposes of this report we are positioning 'digital footprint' data as a subset of the wider group of 'novel' data sources. The composition of this wider group was considered within the 'New Data for Understanding the Human Condition' report,[7] which identified six categories of data sources potentially available for research purposes. All of which are routinely generated, but can potentially be accessed and repurposed for a secondary use:

Category A: Data stemming from the transactions of government, for example, tax and social security systems.

Category B: Data describing official registration or licensing requirements.

Category C: Commercial transactions made by individuals and organisations.

Category D: Internet data, deriving from search and social networking activities.

Category E: Tracking data, monitoring the movement of individuals or physical objects subject to movement by humans.

Category F: Image data, particularly aerial and satellite images but including land-based video images.

The description of Category F above, and the examples provided in the original report, make clear that this category is emphasising spatially indexed records relating to the built and natural environment. We would consider that this category extends beyond 'image data' to also include those collected by sensors in monitoring stations (e.g. temperature, or air pollution records) or modelled from spatial sources (e.g. noise exposure modelled from traffic count records). Other forms of 'image data' – e.g. personal photos – will be considered as part of Category D. As such, these data sources are distinct from all forms of self-reported data including the collection and assaying of biological samples.[8] However, this binary distinction between self-reported data and novel data is also being impacted by the new possibilities inherent in digital devices. For example, *Ecological Momentary Assessment (EMA)* data collection – where participants are invited to provide data on symptoms, affect

[7] Entwisle B, Elias P. New data for understanding the human condition: International perspectives. 2013. Available at: http://www.oecd.org/sti/inno/new-data-for-understanding-the-human-condition.pdf
[8] This distinction is not as clear cut as it may at first seem, as there is potential to link to sources of self-report data (e.g. those provided to different surveys, such as the Pa) or to link to routinely generated assayed samples (e.g. patient test results from pathology laboratories).

and behaviours at a point of time close to experience and at many events or time points[9] - utilises digital devices as the channel to collect self-report data which can also be augmented through the sensor and other data simultaneously recorded by the device itself. For some time, researchers have used smartphone-based EMA for capturing of health-related self-report data in free-living conditions (e.g. Burke et al 2017[10]). EMA approaches prompt participants to answer questions about specific aspects of their health and behaviour several times a day as they go about their normal lives. While these methods can capture high temporal density data, over extended periods they can be become disruptive and lead to high levels of participant burden, which may be particularly problematic in longitudinal cohort studies in which increased burden is a concern for participant engagement.[11] For these reasons, this novel form of data are also considered part of the study's 'digital footprint' strategy and hence included within the scope of this report.

2. The Avon Longitudinal Study of Parents and Children (ALSPAC)

2.1 Study design and sample characteristics

ALSPAC is a multigenerational prospective birth cohort study. ALSPAC recruited pregnant women resident in and around the City of Bristol (South-West UK) and due to deliver between 1st April 1991 and 31st December 1992. There were an initial 14,541 enrolled pregnancies comprising 14,676 foetuses (for these at least one questionnaire has been returned or a "Children in Focus" clinic had been attended by 19/07/99). These pregnancies resulted in 14,062 live births and 13,988 children alive at 1 year. From age seven attempts were made to recruit addition cases who were eligible under the original sample definition.[12,13] By age 24 an additional 913 index children had enrolled. The total sample size for analyses using any data collected after the age of seven is therefore 15,454 pregnancies, resulting in 15,589 foetuses. Of these 14,901 were alive at 1 year of age. Of these, 14,775 were live births and 14,701 were alive at 1 year of age.[14] The cohort has been followed intensively from birth through self- completed questionnaires and attending clinical assessment visits.

ALSPAC has built a rich resource of phenotypic and genetic information relating to multiple genetic, epigenetic, biological, psychological, social, and other environmental exposures and

[9] Moskowitz DS, Young SN. Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. Journal of Psychiatry and Neuroscience. 2006 Jan;31(1):13.

[10] Burke, L. E., et al. (2017). Ecological Momentary Assessment in Behavioral Research: Addressing Technological and Human Participant Challenges. *Journal of Medical Internet Research*, *19*(3), e77. doi:10.2196/jmir.7138.

[11] Lucas, P., Allnock, D., & Jessiman, T. (2013). How are European birth-cohort studies engaging and consulting with young cohort members? BMC Medical Research Methodology, **13(**56).

[12] 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.

[13] Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, Henderson J, Macleod J, Molloy L, Ness A, Ring S. Cohort profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. International journal of epidemiology. 2012 Apr 16;42(1):97-110.

[14] 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.

outcomes. The ALSPAC Web site hosts a data dictionary that describes the available data[15] and further information can be found via the CLOSER Discovery platform.[16]

2.2 The ALSPAC 'Original Cohort Advisory Panel' (OCAP)

The ALSPAC Original Cohort Advisory Panel (OCAP) is a standing committee representing the interests of the index participants of the ALSPAC study. OCAP currently consists of 37 participants (of which 19 participated during 2019) and meets in person approximately every two months.

2.3 Ethics

This report summarises existing research evidence. All the participant consultations reported here were conducted within ALSPAC's ethical framework and with approval of the ALSPAC Ethics and Law Committee (a University of Bristol faculty research committee). Verbatim participant quotes are included with the consent of the participant.

3. Participant views

We will firstly summarise the evidence on the understanding and acceptability of ALSPAC's collection of data via linking to government held health and social records (categories A and to some extent B), given that participant and public views on the use of these data have been reported extensively elsewhere. Secondly, we will synthesise evidence relating to participant understanding and acceptability of digital footprint data collection strategies (categories C, D, E, F) from a range of existing participant consultation exercises.

3.1 Views about consenting to data linkage

Participants views regarding linkage to their official government health and social records were sought in 2011 using qualitative studies involving 55 participants with a range of study participation backgrounds.[17] The interview data suggest that acceptability is influenced by considerations around the social sensitivity of the research question; the need for research to focus on population level patterns rather than have an individual focus; and, the potential for records to deliver of tangible research benefits that are in the public interest. Controlling the identifiability of records in the research process was seen as necessary by most to ensure acceptability; although some participants raised questions about the effectiveness of anonymisation processes and others did not view effective anonymisation as making consent unnecessary. The need for anonymisation was linked to the sensitivity of the data/research question and the potential for stigmatisation or 'labelling' at a personal level and was also linked to the use of information about where people lived (e.g. the use of postcode data in a research project). This was related to notions of ownership of personal information and etiquette around asking permission for secondary use.

[15] https://www.bristol.ac.uk/alspac/researchers/our-data/
[16] https://discovery.closer.ac.uk/
[17] Audrey S, Brown L, Campbell R, Boyd A, Macleod J. Young people's views about consenting to data linkage: findings from the PEARL qualitative study. BMC medical research methodology. 2016 Dec;16(1):34.

Despite different consent procedures being explained, participants tended to equate consent with 'opt-in' consent through which participants are 'asked' if their data can be used for a specific study. This may be explained by the tradition of ALSPAC participation being associated with active consent (for example, agreement to undertake physical assessments at a study clinical assessment visit). Participants raising similar concerns came to differing conclusions about whether consent was needed. Views on this changed when presented with different data use/research scenarios and were sometimes inconsistent.

3.2 Views about the purpose and composition of research ethics committees

The qualitative research seeking views on the acceptability of data linkage also tested participant opinions about appropriate safeguards (such as anonymisation and consent) and research governance, including the purpose and composition of research ethics committees.[18] Of the 48 participants who discussed this topic, most (37) had little or no specific knowledge of ethics committees, while some could remember this being discussed within school lessons. Once given basic information about research ethics committees, only three of the 48 participants suggested there was no need for such bodies to scrutinise research.

The key tasks of ethics committees were identified as monitoring the research process and protecting research participants, although the difficulty of balancing the potential to inhibit research against the need to protect research participants was acknowledged. Participants considered it important that those involved in research oversight should have relevant research and professional expertise, yet it was also considered important to represent wider public opinion, and to counter the bias potentially associated with self-selection to committees through a selection process similar to 'jury duty'. These findings suggest that ALSPAC should emphasise the role of oversight committees in the research process.

3.3 Participant views on the 'spectrum of sensitivity' across digital-footprint records

In 2018 focus groups were held which aimed to elicit opinions on the acceptability of data acquisition from different types routinely generated digital-footprint records. Participants were asked to consider the sensitivity of different data sources in general, rather than whether they would be willing to share these with ALSPAC. Participants were asked to rank a set of 20 different types of data (each printed onto a card, see Table 1) into the order of those which they were most willing to share (least sensitive) to least willing to share (most sensitive).

[18] Audrey S, Brown L, Campbell R, Boyd A, Macleod J. Young people's views about the purpose and composition of research ethics committees: findings from the PEARL qualitative study. BMC medical ethics. 2016 Dec;17(1):53.

Table 1: Cards illustrating a range of routinely generated data sources.

Medical Records	Bank Transactions	Online Shopping History	Social Media	Car GPS
Online Dating History	'Click' History	Mobile Phone Use	Electricity Use	Browsing History
Broadband Use	Mobile Phone GPS	Loyalty Card Data	Home Address	Sleep Patterns
Search History	Age, gender, marital status	Physical Activity (exercise)	Cycling Camera Video	Car Speed Records

The data sources focused on 'digital footprint' data, but also contained a number of more traditional data sources as reference points (i.e. 'Home Address', 'Age, gender, marital status etc', and 'Medical Records'). The discussions held while conducting this exercise suggested participant decisions were influenced by the risks associated with sharing, the potential benefits arising from their research use, and their ability to exercise control. Precise location data (e.g. from Global Positioning System [GPS] sensors in phones or cars, detailed online behaviours, and banking and medical records being considered most sensitive: suggesting that data granularity plays a key part in whether data are considered sensitive.

A detailed description of this study and its findings is available in Annex 1.

3.4 Evidence on the acceptability and understanding of linkage to commercial transactional records

Participant views were tested in focus groups to explore attitudes towards sharing commercial transactional records for longitudinal research and to understand which safeguards should be considered to help ensure the acceptability of this. Participants had little awareness of the research value of transactional records. This finding emphasised the importance of understanding purpose in making decisions about data sharing. Participants raised the need for a number of safeguards, such as the importance of confidentiality, mechanisms to ensure third party users are selected, vetted and controlled and ensuring that participants have access to appropriate information, control over their data in the form of granular consent options and an ongoing right to opt-out. Participants expressed a willingness to share these data, although this willingness was accompanied by a range of concerns and queries about what the process would entail and its potential repercussions. The safeguards identified were seen to be important in maintaining the ongoing trust relationship between the group participants and the study.

A detailed description of this study and its findings is available in Annex 2 and has been published elsewhere.[19]

[19] Please note, this paper is undergoing open review at the time of writing this report. Skatova A, Shiells K, Boyd A. Attitudes towards transactional data donation and linkage in a longitudinal population study: evidence from the Avon Longitudinal Study of Parents and Children. Wellcome Open Research. 2019 Dec 3;4(192):192.

3.5 Evidence on the collection of spatially indexed and location records

ALSPAC systematically records participant address details in order to administer study activities and as a means to conduct spatially informed research. ALSPAC sought the views of the OCAP committee on the understanding and acceptability of using spatially indexed data in ALSPAC research and whether this could be extended to include new address sources (e.g. participants school address) and detailed location/movement data from GPS enabled sensors (such as those found in smart phones and some exercise devices such as 'Strava' and Fitbit). Views were tested using four hypothetical research scenarios that described sharing approximate location (e.g. 1 km2 area), specific locations (e.g. home or school addresses) and exact location (e.g. GPS tracking). OCAP viewed location-based research as broadly acceptable; but expressed preferences for the processing of the location data to be restricted to trusted study staff. Concerns were raised regarding GPS location tracing data, although there was recognition that many individuals were routinely tracked in this way by the commercial developers of device operating systems and applications. Participants expected that confidentiality risks introduced by spatial research are controlled, with the ideal being that all processing of identifiable location data should be conducted by study data managers. However, where this is not possible – such as where particular expertise or equipment is required – then equivalent measures should be deployed (e.g. masking the identifiers of study participants through purposefully including selected addresses of non-study participants). There is a need for transparency in research use of location-based research and for opt-out mechanisms. Further risk assessments and participant consultation is required before precise location tracking data collection is deployed: it is likely that the necessary safeguards needed to ensure acceptability will need co-designing with participants and experts in disclosure risk.

A detailed description of this study and its findings is available in Annex 3 and has been published elsewhere.[20]

3.6 Evidence on the collection of Social Media records

Two focus groups explored participant views towards the acceptability and necessary safeguards needed to support the use of social media data in research. The groups included ALSPAC young people (N=9) and parents (N=5). Participants discussed a range of social media platforms, but were informed that access would only be to their 'visible' information (i.e. what the public or their permitted friends and family could see) and that any access would be made on the basis of opt-in consent.[21] Participants were accepting of ALSPAC's use of their social media records, except private channel information and information about third-parties. This acceptance was based on trust in the study, belief in the purpose for which the data would be used and reassurance about ALSPACs data management processes maintaining their confidentiality. Participant views did not tend to vary across the

[20] Boyd A, Thomas R, Hansell AL, Gulliver J, Hicks LM, Griggs R, Vande Hey J, Taylor CM, Morris T, Golding J, Doerner R, Fecht D, Henderson J, Lawlor DA, Timpson NJ, Macleod J. Data Resource Profile: The ALSPAC birth cohort as a platform to study the relationship of environment and health and social factors. *Int J Epidemiol*. 48:4, 2019. https://doi.org/10.1093/ije/dyz063

[21] The researchers adopted this approach since opt-in consent is a pre-requisite for linking to social media records given that this will need either participants to provide their account details (e.g. their Twitter handle) or to activate data sharing agreements from within their social media account.

different social media platforms. There were some minor differences between the generations on their views of social media and how their social media data could be used for research. These differences may be influenced in the way different generations use social media, and by differing conceptions of risk and how systems worked.

A detailed description of this study and its findings is available in Annex 4.

3.7 Evidence on an Ecological Momentary Assessment (EMA) study of alcohol consumption.

OCAP's views were sought on the understanding and acceptability of collecting EMA data through an ALSPAC issued smartwatch. OCAP focused their discussion around the initial proposed use of the device to collect information on alcoholic drinking behaviours. OCAP members were broadly supportive of this proposal. There was limited discussion of risks and the perceived issues related to intrusion (annoyance), participant burden and the potential for error/poor data quality. The scope of the raised concerns is likely to result from this being a consented ALSPAC-centric exercise where the study controlled all aspects of the process, including the device and data. Furthermore, the proposal did not extend to using device-based sensors (e.g. GPS location). OCAP's guidance suggests the need to ensure such devices and software are easy to use and minimise participant burden. Ideally, participants would rather use their own device rather than a study issued one.

A detailed description of this study and its findings is available in Annex 5.

4. Ethical Issues

The use of available 'novel data' in UK longitudinal research is not new: Doll & Hill linked mortality records to a cohort of British doctors in 1951,[22] and Acheson built the Oxford Record Linkage Study resource by linking hospital admission statistics to mortality records in the 1960s.[23] Furthermore, LPS have often had somewhat of a magpie approach to data collection, given that their broad purpose is to build a research databank and that it may not always be clear as to which data will be relevant for future research studies. Therefore, it may be reasonable to consider that the use of digital footprint data is not a fundamentally new principle, but rather a continuation of existing study objectives through an innovative use of data. However, this view needs to be considered against evolving expectations regarding research ethics, wider societal changes in the use of individuals' data and a realisation there is a 'social licence'[24] needed to ensure acceptable and sustainable data-intensive research. We consider the ethical issues arising from the participant consultations through the lens of a set of questions.

Is there anything meaningfully different between digital footprint research and other longitudinal research activity?

[22] Doll R, Hill AB. The mortality of doctors in relation to their smoking habits. British medical journal. 1954 Jun 26;1(4877):1451.

[23] Acheson ED. Oxford record linkage study: a central file of morbidity and mortality records for a pilot population. British journal of preventive & social medicine. 1964 Jan;18(1):8.

[24] 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.

ALSPAC operates within a data usage framework it has established with its participants (see section five below), the frameworks and expectations set out by its funders (for example the 'Bona Fide' principles for data sharing set out by the Medical Research Council), expectations of good-practice within the wider academic community (e.g. making data discoverable and accessible through managed processes), and those imposed by ethical review bodies, in legislation and in regulatory codes of practice. The bounds in which ALSPAC operate allow for innovation within a broad range of operational parameters, but any profound change in purpose would need to be communicated to participants and would likely require consent. There is no such perceived change in purpose through collecting and using digital footprint records to inform epidemiological/social science investigations aiming to improve the public good.

Some challenges faced when using digital footprint records are permutations on the old, for example new considerations in maintaining confidentiality reflect the existing challenge of using rich individual level data which is inherently identifiable. Here the 'new' considerations relate to the complexity and variety in data, for example de-identification mechanisms used to de-identify free-text data will be different to those used for image data. A pragmatic approach to addressing these risks would be to restrict access to the raw, granular data to study data managers and to only permit researchers to access derived variables. Our evidence suggests this approach would be acceptable (in terms of addressing confidentiality risk) to participants, yet would impose resource issues on the studies and may require specialist expertise.

Other challenges are new. For example, some sources – particularly social media records and transactional records – will include information on identifiable third parties such as friends or other household members. Our findings suggest that many participants would not find this acceptable as this would represent a breach of trust between themselves (who would be enabling the use of data) and their friends and/or household members (who would not have visibility or control over this). It is not clear if anonymisation approaches would resolve this. Where possible these third-party data should not be collected or only collected at an aggregate level (e.g. the number of 'friends' an individual has, or the level of engagement with a particular social media post), although this may limit some research options.

Does digital footprint research expose participants to meaningful new risk?

Many of the perceived risks identified by participants are based on fears relating to data which would not be collected (e.g. the collection of bank account numbers could increase fraud risks; whereas these would be anonymised prior to study access), or relate to misconceptions of how digital footprint research would operate in practice (e.g. that the research could impact on credit ratings). Participants did recognise that many of the perceived risks relating to the use of these data in longitudinal research already exist due to the systematic use of these data by commercial companies who are monetising digital footprint data as a new commodity.

The concentration of increasingly diverse and sensitive linked information about individuals in one setting – i.e. a study databank – could reasonably be seen to increase risk, as one system is used to protect myriad sensitive data. This reinforces the existing recognition that study databanks need not only to be secure but require frequent auditing[25] and need to be subject to continuous improvement (e.g. operating to ISO27001 Information Security standards). It also strengthens the case for keeping an 'air gap' between the systems containing personal identifiers (the study contact database) and the systems containing pseudonymised research data. Separated systems such as these are recognised in regulatory codes of practice (i.e. are part of Data Protection minimisation principles) and can be enforced at a study/institution level or through using infrastructure providers such as UK Secure eResearch Platform which hosts data but does not contain direct personal identifiers.

Is digital footprint data collection and research perceived differently?

There was a trend across the different focus groups that the purpose for using digital footprint data in research was not immediately clear, and that acceptability was tied to the data use and research having a clear purpose and confidence that the proposed data use would lead to public benefits. It is paramount that studies effectively communicate the benefits that can be realised through using these data sources and it would be prudent to seek insights on how to best achieve this from initiatives such as Understanding Patient Data.[26]

Participants strongly suggested that some digital footprint records were highly sensitive, but also suggested that some more traditional data sources – such as linked health records – were at least as sensitive if not more so. Many agreed that they would be prepared to share their records, even those which were sensitive, with ALSPAC. However, this evidence suggests that to maximise acceptability some information (e.g. social media data on third parties) should be excluded from data extracts and that participants should be given granular consent/objection options.

What are the new responsibilities for the study resulting from digital footprint strategies?

ALSPAC has an existing participant safeguarding and reporting procedure, particularly in relation to our statutory duties relating to situations where a child is suspected to be at risk of harm. It should be considered whether these procedures need updating to reflect risks identified within digital footprint records. Although it must be recognised that most of the data are unlikely to be systematically studied by a data manager and that data will not be accessed or processed in a timely manner. Aligned with this, it has been recognised in some ethical frameworks[27] that there is an ethical duty to protecting staff members from harms resulting from research using digital footprint records. This could relate to existing risks and safeguards – for example ensuring procedures are in place in the event of staff spontaneously re-identifying a participant they know and then having to manage the

[25] Auditing in this sense is taken to include regular assessments of system vulnerabilities, known as 'penetration testing', and also assessments as to the rigour and suitability of an organisations policy and practice (for example, ISO27001 audit assessments).

[26] See https://understandingpatientdata.org.uk/

[27] franzke, aline shakti, Bechmann, Anja, Zimmer, Michael, Ess, Charles and the Association of Internet Researchers (2020). Internet Research: Ethical Guidelines 3.0. https://aoir.org/reports/ethics3.pdf

confidentiality of this information – to new risks, such as encountering disturbing content in social media posts. These areas have not received much attention within the longitudinal community and would benefit from cross-cohort consideration.

ALSPAC is responsible for providing high-quality data to its researchers and sufficient documentation to allow researchers to understand the data processing history. There will be a challenge for data managers to document these increasingly complex data and their provenance, particularly where access to the underlying raw data is restricted.

5 ALSPAC's Digital Footprint strategy

5.1 ALSPAC's Digital Footprint data strategy

There is funding in place in ALSPAC to establish linkages to social media records and active programs considering linkage to transactional records (e.g. store loyalty cards, banking records), making further use of phone and smart devices (e.g. EMA studies) and through the use of sensors (e.g. modelling of pollution, green space and noise data collected via remote sensors and neighbourhood monitors and linked to participants through residential address). ALSPAC have established a Digital Footprint working group to coordinate the development of an overarching strategy, coordinated participant communications and safeguards and to share insights into acceptability and good practice.

ALSPAC's approach to undertaking these linkages is to:

1) test views on participant understanding and acceptability;
2) to work with domain experts and data owners to understand which data are available, to identify potential data linkage and extraction mechanisms,[28] and to interrogate the quality and bias in each data source;
3) to conduct risk assessments as mandated in Data Protection legislation;
4) to build data linkage, processing, and transformation pipelines incorporating sufficient controls to meet participant expectations, to deliver high-quality and fully documented data, and to ensure ethico-legal compliance;
5) to gain relevant ethico-legal approvals for these processing activities;
6) to explain our proposed use of these data to participants and to collect consent where practicable or to provide mechanisms for objecting where not;
7) to communicate the availability of these data to the (legitimate) research community;
8) Disseminating findings in order to promote the resource and to build participant engagement in the study.

[28] This will include technical and legal considerations to establish ownership rights, Data Controller/Processor status, data use conditions and to identify the mechanisms by which these conditions can extend through the data processing stages and onto third-party researchers during the analysis stages (e.g. onward sharing contracts, mechanisms for upholding responsibilities such as auditing).

It will be essential to clearly and regularly communicate this intended use of data to participants and to build a case as to why these data can be used safely to realise benefits for the public good. Key messages will be that this strategy reflects changing opportunities brought about by an increasingly digitised society and not a change in study principles or aims, that ALSPAC's strategy aligns with wider aims for data-driven policy development and service provision (e.g. that the NHS has similar ambitions aiming to use data to bring public goods), and that participants retain control even where approaches are not based on explicit consent.

5.2 Impact of participant views on the ALSPAC strategy

Participant consultation has resulted in a series of meaningful impacts on ALSPAC's policies and practice. Highlights relevant to our digital footprint strategy are described below.

5.2.1 'Our Commitment to you'

Our evidence suggests that any requests for participant permissions will need to be balanced by formal commitments by the study as to how the data will be used. To help address this, ALSPAC have developed a formal set of study principles (see Table 2) that are regularly communicated to participants in the form of a signed declaration by the study Principal Investigator. These principles emphasise autonomy, the role of ethical oversight, the principle of confidentiality and that the study's research aims to deliver public benefits and will not be used for profit. These principles are backed with relevant study policies and operating frameworks (e.g. a participant withdrawal policy) which have been subject to participant oversight and ethical approvals.

Table 2: the 'Our Commitment to You' principles

- Taking part in the project is voluntary and you are free to withdraw at any time without giving a reason.
- You will not be identified from the research - researchers do not see your name with your information – they just see your barcode ID number.
- Every research project is checked to make sure it meets the highest scientific and ethical standards.
- In the same way as a doctor who treats you is bound to keep your information confidential, Children of the 90s, and all the researchers we work with, are bound to keep your information confidential.
- There are independent experts whose job it is to look at what we do and how we do it to make sure your rights are protected.
- We do not do research with the aim of commercial gain - all our research aims to benefit society and is not for profit.

ALSPAC have a dedicated communications team who work with participant advisors to produce clear and transparent study information aiming to ensure there are 'no surprises' in how the study uses participant data. Study information is provided through: formal 'fair processing' websites,[29] newsletters[30] and information leaflets;[31] our study social media channels;[32] and participant events. New information is provided in response to changing context of data use, for example a recent newsletter set out information in response to new Data Protection regulations.[33] The variation in participant expectations for detailed information has led to a strategy where detailed and technical information is provided on a 'layered' basis for those who wish to know.[34]

5.2.2 Approaches to 'consent'

All participation in ALSPAC is voluntary and providing data, or involvement in any given part of the study is discretionary.

The 'consent' strategy for linkage to official health and administrative records seeks to balance participant views on the importance of autonomy and etiquette with the potential for non-response to consent campaigns to bias research findings and to marginalise some vulnerable population groups. ALSPAC has identified there are systematic differences between those active in the study and those lost to attrition[35] and is utilising data linkage as a means to address this differential response, to correct for potential bias[36] and as a means to include vulnerable and marginalised sub-groups in our research.[37] To mitigate the impact of this differential attrition ALSPAC have adopted an 'opt-out' approach for linkage to these records. Participants are given clear information about the studies intended use of their records, and have the right and means to object. Granular decision choices are given to reflect some participants finding different data sources more or less acceptable. Regular reminders are given about this approach to data linkage in newsletters and other study communications. Opt-in consent is taken when practicable (e.g. when a participant attends a research clinic) and for our next generation of participants where consent is sought from the index participant at the point of enrolling their next generation child.

In contrast, the consent design for digital footprint records is heavily influenced by pragmatics, where most digital footprint linkage requiring active participant involvement (e.g. participants need to actively tell the study what their Twitter account handle is, or participants need to actively authorise the release of data from within a system, or to

[29] http://www.bristol.ac.uk/alspac/participants/privacy/
[30] http://www.bristol.ac.uk/alspac/participants/newsletters-leaflets/
[31] http://www.bristol.ac.uk/alspac/participants/using-your-records/
[32] See: https://www.facebook.com/childrenofthe90s; https://twitter.com/CO90s https://www.youtube.com/user/children90s; https://www.instagram.com/children_of_the_90s/
[33] http://www.bristol.ac.uk/media-library/sites/alspac/documents/newsletters/CO90s-familynewsletter-2018.pdf
[34] This approach is based on good-practice guidance issued by the Information Commissioner's Office and is designed to meet the Data Protection Act 2018 'right to be informed'. Available from: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-be-informed/
[35] Ibid 10.
[36] 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 Apr 8;44(3):937-45.
[37] Teyhan A, Boyd A, Wijedasa D, Macleod J. Early life adversity, contact with children's social care services and educational outcomes at age 16 years: UK birth cohort study with linkage to national administrative records. BMJ open. 2019 Oct 1;9(10):e030213.

participate in an EMA exercise). Therefore, digital footprint linkage and EMA studies will likely be based on opt-in consent. The information materials provided will be tailored to address areas of uncertainty raised at the focus groups and the safeguards participants considered necessary.

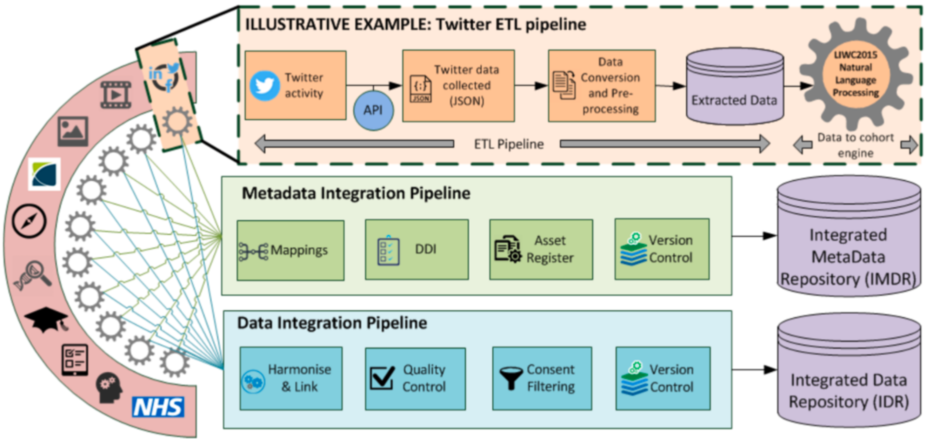
This contrasting approach is being explained to participants through our regular communications:

> "**Sometimes we ask you to 'opt in' for your consent to a new project but sometimes we let you know about our plans and give you the right to 'opt out'.** We are approaching things in this way because we want to make sure that the findings from Children of the 90s, and therefore the evidence that informs the NHS and government decisions is fair and inclusive to all." ALSPAC 2019 participant newsletter.[38]

5.2.3 The data linkage and digital footprint data pipeline

Participant feedback has stressed the importance of maintaining confidentiality and ensuring study control of the data. By necessity all linked data are identifiable and considered Personal Information in Data Protection legislation (even when pseudonymised). The Project to Enhance ALSPAC through Record Linkage (PEARL) have established a data extraction and processing pipeline (see Figure 1) which is built to 'Data Safe Haven' principles.[39] A central principle in this approach is that access to sensitive identifiable information is restricted to study Data Managers.

Figure 1: The PEARL data processing pipeline


ILLUSTRATIVE EXAMPLE: Twitter ETL pipeline
Twitter activity
API
Twitter data collected (JSON)
Data Conversion and Pre-processing
Extracted Data
LIWC2015 Natural Language Processing
ETL Pipeline
Data to cohort engine
Metadata Integration Pipeline
Mappings
DDI
Asset Register
Version Control
Integrated MetaData Repository (IMDR)
Data Integration Pipeline
Harmonise & Link
Quality Control
Consent Filtering
Version Control
Integrated Data Repository (IDR)
NHS

This data model comprises three main stages:

1) **Extract, Transform, Load Stage (with a break-out example shown in Figure 1 using the Twitter linkage as an example):** each data type (e.g. Twitter, Store loyalty cards, NO2 air pollution models) is linked to and the

[38] http://www.bristol.ac.uk/media-library/sites/alspac/documents/newsletters/CO90s-familynewsletter-2019.pdf

[39] Burton PR, Murtagh MJ, Boyd A, Williams JB, Dove ES, Wallace SE, Tasse AM, Little J, Chisholm RL, Gaye A, Hveem K. Data Safe Havens in health research and healthcare. Bioinformatics. 2015 Jun 25;31(20):3241-

maximum amount of acceptable information is captured. The data are stored within the PEARL secure servers in their raw form.

2) **Processing engine:** software routines convert the raw identifiable data into derived and de-identified values. E.g. in the Twitter example, the identifiable and sensitive raw Tweets are processed into derived sentiment scores using the validated LIWC2015 natural language processing tool.
3) **Integration pipelines:** the derived values are integrated into ALSPAC's self-reported and other data and separately into the metadata catalogues.

This approach maximises the potential for future use by allowing study Data Managers (as trusted individuals) to curate rich, identifiable data and to de-identify and minimise these to each future research potential. In keeping with participant expectations and the Data Safe Haven model, all systems, policies and procedures are independently audited and certified to the ISO27001 Information Security standard and staff are subject to vetting, training and penalties for misuse of data. The most sensitive and complex linked records will only be available via secure research servers which do not permit the physical sharing of data, but allow remote access to conduct research under secure and audited conditions.

5.2.4 Data Access Policies

The ALSPAC researcher Data Access Policy[40] has been enhanced with additional safeguards (e.g. the 'split stage protocol' which introduces safeguards where identifiable data such as residential address information are needed in the research process). The key aspects of the policy have been incorporated into an enhanced and legally binding Data Access Agreement[41], which also contains mechanisms to 'onwardly share' requirements of the original data owners in projects using linked data.

5.2.5 Ethics

ALSPAC's ethics framework includes the OCAP committee who provide guidance and help develop study information materials. Projects – including all new data collection activities – are then required to seek University of Bristol faculty ethics approval. ALSPAC's requests for faculty ethics approval are considered by the ALSPAC Ethics & Law Committee (ALEC). This committee aims to have a membership comprising both study participants and expert members (sometimes both). In line with participant feedback (and wider expectations of good practice), the ALEC committee brings a participant as well as expert voice to ALSPAC's approval process. The role of the committee in protecting participant interests is actively promoted to participants.[42]

6. Conclusions

[40] Available from: http://www.bristol.ac.uk/media-library/sites/alspac/documents/researchers/data-access/ALSPAC_Access_Policy.pdf

[41] Available from: http://www.bristol.ac.uk/media-library/sites/alspac/documents/researchers/data-access/alspac_data_access_agreement.pdf

[42] Available from: http://www.bristol.ac.uk/media-library/sites/alspac/documents/newsletters/CO90s-familynewsletter-2019.pdf

ALSPAC have identified that digital footprint records offer new research utility and are starting to develop the mechanisms for incorporating these data into the study databank. It is imperative that this novel use of participant data is understood and acceptable to participants in order to maintain trust in the wider study.

Qualitative approaches to understanding participant views on the mechanisms by which this happens and the safeguards needed are key to ensuring this acceptability. The evidence collected to date suggests that such use of novel data can be acceptable if the value of the data and the likely benefits can be communicated clearly, if confidentiality can be maintained and if participants are offered mechanisms to control the use of their data. A successful digital footprint strategy will depend on 'team data science' approaches which bring together expertise in study management, data and informatics, participant communication and engagement, research expertise along with an active and meaningful public/participant role. To sustain this strategy, ALSPAC will need to continue to listen to participant views and to be responsive to emerging issues.

Annex 1: Participant views on the 'spectrum of sensitivity' across digital-footprint records

A1.1 Introduction

Focus groups were held that aimed to elicit opinions on the acceptability of data acquisition from different types of routinely generated digital-footprint records. The overall aim of the focus groups was to explore the understanding of, attitudes towards, and safeguards needed to use transactional records - e.g. store loyalty cards and banking data for longitudinal research (we discuss this in depth later in this report). However, the initial focus group (Focus Group 1) exercise considered the use of many diverse types of digital-footprint data. The exercise aimed to understand reasonable expectations about the sensitivity of different data sources in relation to their access and use across society, rather than in relation to data use in longitudinal research. The following results have not been published.

A1.2 Methods

Participants were invited to attend a series of three linked focus groups which were held during 2019, when participants were approximately 27-28 years old. During the first focus group an exercise was conducted to test views relating to the sensitivity of 20 different types of data. A very high-level name for each data source was printed onto a deck of 20 laminated cards (See Table 1). The data sources focused on 'digital footprint' data, but also contained a number of more traditional data sources as reference points (i.e. 'Home Address', 'Age, gender, marital status etc', and 'Medical Records'). Participants were given an introduction to the task and an overview of what the datasets meant and asked to rank the data sets in order of sensitivity. In this context, the sensitivity scale was summarised to participants as ranking the cards in order from which they were most willing to share (least sensitive) to the least willing to share (most sensitive).

Table 1: Cards illustrating a range of routinely generated data sources.

Medical Records	Bank Transactions	Online Shopping History	Social Media	Car GPS
Online Dating History	'Click' History	Mobile Phone Use	Electricity Use	Browsing History
Broadband Use	Mobile Phone GPS	Loyalty Card Data	Home Address	Sleep Patterns
Search History	Age, gender, marital status	Physical Activity (exercise)	Cycling Camera Video	Car Speed Records

Focus group 1 [FG1] was run twice, with 10 participants attending the first instance [FG1a] and 6 different participants attending the second [FG1b]. For this task, the participants of FG1a and FG1b were initially split into three sub-groups. The sub-groups each sat around their own table and had their own deck of data source cards (each set was randomly sorted). Each sub-group was asked to independently rank the data sets in order of sensitivity. Following this, the sub-groups were convened into a whole group exercise, where all participants were asked to collaboratively agree a consensus ranking. The order in which the data source cards were ranked was recorded.

The results have been thematically coded using both the sub-group and consensus outcomes. The consensus outcomes have been given greater emphasis given that these outcomes were determined following additional group discussion and the participants making further requests for information from the academics.

A1.3 Results: participants consideration on sensitivities

The following results are currently unpublished.

The thematic coding identified the following themes in the discussions around sensitivity: the sensitivity of a data source linked to the risk of sharing; weighing the balance between risks and benefits; sensitivity in relation to personal choice and control; and, finally sensitivity in relation to the granularity of information within the data source. We summarise evidence within these themes in turn.

Sensitivity linked to risks of sharing data

When ranking the various types of digital footprint data according to sensitivity, participants considered the risks that may be associated with the different data types. Age, gender and marital status were largely considered to be low risk, as participants felt this information could easily be deciphered. Participants also felt that there were little risks associated with physical exercise and sleep data.

Digital footprint data were considered more sensitive if they revealed location, which had implications for spatial data collection. This included linking to GPS sensors within different devices:

> *'I would say it's because they know your routine more than anything.'* (FG1b, Group discussion)

> *'Yeah, where you shop, where you work, where you go out.'* (FG1b, Group discussion)

Participants also discussed a number of possible risks associated with sharing bank transactions. For some salary was considered very sensitive, but for others it was information they would happily share. Concerns around sharing bank transactions were more frequently related to financial implications:

> *'Could there be a chance that that might impact the deals you get from your bank maybe?'* (FG1b, Group discussion)

> *'Will it affect things like your credit history and stuff?'* (FG1b, Group discussion)

Participants also feared the ways in which data could be manipulated and used against them. For instance, cycling camera video:

> *'You're in an accident with a car, a motorist and generally in my experience they're both as bad as each other but you're the person with the power, you can cut and edit it. So it makes the person who was in the car look worse.*' (FG1b, Group discussion)

There was also a risk of embarrassment associated in particular with both medical records and online dating records:

> *'I think that'd be quite embarrassing if you found out that somebody was using, you know, knowing what, I don't know, prescriptions you've had in your life or any illnesses you've had.'* (FG1a, Group discussion)

> *'If I had a dating app and somebody can access who I'd like or who I might message and what would I be saying, I'd be mortified…'* (FG1a, Group discussion)

Data containing information relating to others was also considered sensitive if there is lack of consent to share or use this data:

> *'It's all about permission […]. Because if you post that out there and they want to take it down and you haven't got evidence that you've got permission from them then it's a legal battle really from there.'* (FG1b, Group discussion)

One participant emphasised how sensitivity is associated with who is sharing data, and in particular, whether they are aware of the risks. For instance, children:

> *'So like social media as adults I don't think it's that big a deal but obviously with kids you know they can share god knows what with who.'* (FG1b, Group discussion)

For one participant, sharing personal data was considered high risk to the extent that he would prefer to pay to avoid sharing data:

> *'I'd much rather pay a small fee for a lot of the services that I use, and not share my information. Like, not because of the fact I'm worried about anything being found out about me right now, but if I were in a predicament, I don't really want to have all of that data out there about me.'* (FG1a, Group discussion)

Sensitivity versus benefits of sharing data

When making decisions about the sensitivity of various types of digital footprint data, participants also considered the possible benefits of sharing data and whether these outweighed the risks. This ranged from finding useful products from targeted ads using online shopping data; better tariffs based on electricity and broadband use; receiving support on Twitter as a result of sharing personal stories, and the importance of cycling camera video in providing evidence in case of accidents.

In addition to personal benefits, participants also made reference to the ways in which sharing digital footprint data could benefit society, such as raising awareness about certain health conditions:

'I had [a rare form of disease] when I was a child, […], so if you don't know someone that's had it, you know nothing about it. But having things like social media and people being able to share it, you can sort of see these things and you can find out about symptoms.'* (FG1b, Group discussion, *disease type suppressed for disclosure control reasons)

Similarly, even though medical records were ranked the most sensitive of digital footprint data types, participants reflected on how they are willing to share these types of data in certain circumstances, such as ALSPAC research:

'I mean we've got medical records as the worst one we'd share but we've all shared and done tests here and had stuff poked and prodded and filmed.' (FG1b, Group discussion)

Sensitivity in relation to personal choice to share data

Participants across both focus groups discussed the ways in which sensitivity was associated with the degree of choice and control they had over sharing the different types of digital footprint data. For instance, various different forms of datasets shared online, such as social media posts (with the exception of private messages) and online shopping history, were associated with personal choice and therefore considered to be less sensitive:

'If you're putting your data online, like anywhere, you have to be kind of prepared for the fact that you are going to lose control of that data immediately, so if you don't want to lose control of that data, that's when you have to kind of, keep it, you know, close to the chest and all that.' (FG1a, Group discussion)

However, search history was considered to be more sensitive in this instance:

'Whereas my search history is like, everything and anything I might ever need, think, wonder, Google, but at least social media I can kind of, make the choice of what I'm saying.' (FG1a, Group discussion)

Sensitivity was also related to whether participants had a choice with whom to share their data and whether they were trustworthy:

'We'd be most offended if we found out someone was using it I suppose, or yes, collecting it without our consent.' (FG1a, Group discussion)

For example, participants felt safer sharing data with ALSPAC as they knew they could trust them to not misuse data, and were able to approach them at any time to delete or stop collecting data:

'But what you don't have with the companies I suppose [unlike ALSPAC], is, well, I don't feel like I have that power, like I don't feel like I could approach Google and be like – right, delete everything you have about me.' (FG1a, Group discussion)

Sensitivity in relation to data granularity

There were a number of remarks made in relation to how sensitivity of data was dependent on the granularity of data that could be accessed:

'I don't think I mind it if it's, if it's anonymous […] because it can't be attributed back to you in any way. So long as there's no way of linking it back to you, I don't really see the harm in that I don't think. If it's just a number.' (FG1a, Group discussion)

For instance, with mobile phone records, one participant was happy to share the number of phone calls and text messages she sent every day, but not the numbers and content of text messages:

> *'So if it's your phone numbers and the conversations that you're having then I wouldn't be ok with that, but if it's just, I don't know how often I send a text or, I don't know, how long I spend on the phone that wouldn't be an issue.'* (FG1b, Table 3)

A similar decision was made around electricity and broadband use:

> *'Well we put phone use, electricity use and broadband use [lower in sensitivity], because it's just how much you're using as opposed to what you're using.'*(FG1b Group discussion)

Likewise with loyalty cards:

> '*I think the thing about all of this is it depends on whether it's connected to you as an individual, so whether people know that it's you or it's collected anonymously. So, whether or not it's just you are making up part of a set of statistics. Because if you've got loyalty card data being shared but its just people seeing shopping habits of all customers in a particular store, they're not necessarily seeing you as an individual associated to a particular person.'* (FG1a, Table 3)

One participant made a similar comment about banking transactions:

> *'It depends on how it's collected and how confidential the data's going to be, how anonymous the data's going to be etc, because like, if you're looking at us as a wide group and you're seeing like, one of us bought flowers on a particular date, that's not really an issue but like, if you're looking at each individual and you're seeing personal transactions that's more confidential.'* (FG1a, Group discussion)

3.1.4 Results: participants ranking of the sensitivities of different data sources

There were distinct patterns of agreement across the sub-groups and consensus groups and between FG1a and FG1b (see Table 2). Medical records and Bank transaction records were generally viewed as being the most sensitive. In the FG1b whole-group consensus exercise the participants could not agree on whether banking transactions were considered sensitive. Whilst one participant felt that this form of digital footprint data should be high on the scale of sensitivity, others disagreed because records would not show what you bought or how much was in your account. There were also disagreements on salary and whether this was considered sensitive. Furthermore, there was confusion about whether sharing this form of data could have impacts on credit history or deals from the bank. No consensus was reached. Basic demographic information (Age, gender, marital status etc) was typically seen as low sensitivity, as were records of physical activity and sleep patterns. Records of patterns in phone and broadband service use were seen as distinct from, and less sensitive than records of detailed service activity (e.g. click history). Although loyalty card data was consistently seen as less sensitive than granular records of phone or internet use despite holding granular purchase history data. There was a difference in opinion between FG1a and FG1b regarding the sensitivity of electricity use: this was a result of a participant in FG1a working in a similar sector and having knowledge of the type of information that is inferred from service use data (e.g. number of household occupants, behaviour information).

Data sources recording precise location – including home address and the use of GPS sensor readings – were consistently considered to be amongst the most sensitive of the data sources.

Table 2: Participant ranking of the sensitivities of different data sources

	Ranking of Sensitivity*							
	FG1a				**FG1b**			
Data Source	**Table 1**	**Table 2**	**Table 3**	**Consensus**	**Table 1**	**Table 2**	**Table 3**	**Consensus**
Medical Records	5	20	20	20	8	20	18	19
Bank Transactions	5	19	19	19	9	19	20	n/a
Online Shopping History	1	13	9	8	10	12	4	8
Social Media	2	8	6	10	12	11	1	11
Car GPS	3	7	9	16	18	14	11	14
Online Dating History	5	9	9	11	11	9	6	11
'Click' History	2	13	2	11	15	12	5	15
Mobile Phone Use	4	11	6	11	14	4	13	9
Electricity Use	2	5	14	5	4	4	1	4
Browsing History	4	13	n/a	11	15	17	10	15
Broadband Use	4	13	14	5	4	4	13	4
Mobile Phone GPS	3	18	14	16	20	15	13	18
Loyalty Card Data	2	6	2	8	6	7	1	6
Home Address	5	12	14	16	7	16	18	10
Sleep Patterns	1	4	6	2	1	1	7	1
Search History	4	13	9	11	15	18	5	17
Age, gender, marital status etc	1	10	1	1	3	10	17	7
Physical Activity (exercise)	1	3	2	2	1	2	7	2
Cycling Camera Video	5	1	2	7	13	8	7	11
Car Speed Records	3	2	13	2	18	3	11	2

* Many groups ranked multiple data sources as being the same level of sensitivity, or clustered sensitivity into groups. The rankings expressed therefore have many tied values. The colour-coding reflects the quintile of sensitivity ranking, with the shading progressing in density as the sensitivity increases.

A1.5 Conclusions

Across the sub-group discussions as well as the larger group discussions, participants reached similar conclusions as to the spectrum of sensitivity across digital footprint data. Medical records were clearly the most sensitive form of data for both groups. However, as one participant explained, participants taking part in these focus groups had agreed to sharing their medical records with ALSPAC in the past, indicating that individuals may be willing to provide access to even the most sensitive forms of data given certain factors, such as clear benefits of donating data, ongoing choice and control over their data, and trust in the organisation using the data through the establishment of appropriate safeguards.

Bank transactions were ranked as the second most sensitive form of digital footprint data in FG1a. However, participants in FG1b could not reach an agreement on their ranking for this form of personal data. Participants were unsure of the granularity of data that could be exposed through access to their bank transactions, which they also discussed in relation to other categories of data, such as mobile phone records. There seemed to be broad agreement that data showing patterns of use (e.g. duration of social media use) was less sensitive than data showing specific itemised use (e.g. the content of social media posts). Data showing precise location were considered sensitive. An individual's understanding of the granularity of data that can be accessed through the various categories is therefore influential on their decision as to the sensitivity of their digital footprint data.

Annex 2: Evidence on the acceptability and understanding of linkage to commercial transactional records

A2.1 Introduction

Transactional records generated during interactions with commercial companies - such as banking records and retail loyalty cards - offer opportunities for longitudinal population studies to capture data on participants' real-world behaviours and interactions. Participant views were tested to explore attitudes towards sharing commercial transactional records for longitudinal research and to understand which safeguards researchers should consider implementing when looking to request transactional data from participants. The findings of this focus group work have been reported in detail elsewhere[43].

A2.2 Methods

Participants were invited to a series of three focus groups, with the first meeting being run twice with different participants – as described in 3.1 above. The groups used semi-structured discussions designed to elicit opinions. Overall, 20 participants attended at least one focus group. In contrast to the activity described in Annex 1, these findings report information from all three focus groups (FG1, FG2, FG3) but only in relation to the use of transactional records. The three sequential focus group model aimed to facilitate more in-depth discussions around the potentially complex topic of using transactional records in longitudinal research and to help encourage an evolution of views. Thematic analysis was used to sort data into overarching themes addressing the research questions.

A2.3 Results

Participants expressed a variety of attitudes towards transactional data linkage, which were associated with safeguards to address concerns. Initially there was confusion amongst many as to how transactional data could be used for research:

> *"What are you guys hoping to achieve by understanding what we're buying, and how is that going to help future generations?" [FG1]*

And how the research use of the data may impact on the services they receive:

> *"Could there be a chance that that might impact the deals you get from your bank maybe?" [FG1]*

> *"Will it affect things like your credit history and stuff?" [FG1]*

[43] Skatova A, Shiells K, Boyd A. Attitudes towards transactional data donation and linkage in a longitudinal population study: evidence from the Avon Longitudinal Study of Parents and Children. *Wellcome Open Res*. 2019;4:192. https://doi.org/10.12688/wellcomeopenres.15557.1

Overall, 'trust' was a major theme, with ALSPAC typically being considered trustworthy, yet this trust not necessarily extending to other third-party users:

> *"The motivations of Children of the 90's, some policy makers and people who are researching or looking into rare diseases that would be, erm, I don't know, it's something about their motivations just seems more legitimate." [FG2]*

But that this could potentially be mitigated through vetting approaches, restricting sharing to trustworthy organisations and ensuring users of the data maintained the standards established by ALSPAC with their participants:

> *"Can we trust them? I guess you would look into it wouldn't you and see what other people have said about them." [FG2]*

> *"I've said people are sceptical or have no trust when it comes to data but not around more trustworthy organisations or organisations you have a trust in history." [FG2]*

> *"So long as they act in the same way that Children of the 90's do then I don't see any problem of having the same level of data that you do."* [FG3]

Significant emphasis was placed on the safeguards which helped generate that level of trust. These are linked to the transparent use of information:

> *"I personally would probably give you all of that [transactional data] if I had a sheet explaining that you were going to do something with it and I was happy with the purpose."* [FG1]

And security controls protecting confidentiality:

> *"That's one of the things I really like about Children of the 90's, they collect all of this data but they keep it anonymous and confidential, always."* [FG1]

> *"You're just a number and you know your shopping habits or your banking habits are just part of a bigger data search. That feels, that feels safer doesn't it?"* [FG2]

> *"I don't think I mind if it it's anonymous. [...] So long as there's no way of linking it back to you. [...] If it's just a number."* [FG1]

While, in general, there was acceptance that participants would share transactional records with ALSPAC, this was not a universally held view:

> *"I don't really think I want people to know where I shop and how often just because, I don't know, it's a bit personal." [FG3]*

And degrees of sensitivity of the data within the record were noted:

"I don't really mind if people know I'm buying shoes or meat and groceries and stuff. But I think the only sensitive, I don't know whether it should be sensitive or a privacy issue, is prescription medicine, contraception maybe." [FG3]

"I think there will be some people that would have a problem with it. Maybe for things like you don't want people to know how much you earn or if you're gambling a lot. So people that have some insecurities maybe don't want other people to know about it." [FG3]

Resulting from this, the importance of consent was discussed as was the ability to change decisions into the future:

"Yeah, I'll probably say yes to anything and everything [...] just as long as I knew what it was and if I did ever want to change my mind, I knew that I was free to do so." [FG2]

"So, I think if we had something we could, like a link or something, we could click on at any point and opt out I think that would be good." [FG3]

Participants advised that consent forms should be structured in such a way as to allow participants to fine tune their consent decisions:

"If you just say transactional data, if someone doesn't really want online stuff within that they will just say no to the whole thing. Whereas they might have been happy for the loyalty cards stuff." [FG2].

A2.4 Conclusions

The focus groups suggest that there is little awareness of the research value of transactional records amongst the public, even engaged and lifelong members of a longitudinal study. This finding is particularly important as this focus group work again emphasised the importance of understanding purpose in making decisions about data sharing. There were a number of safeguards raised which studies should consider when looking to communicate information about transactional record linkage, such as the importance of confidentiality, about how third party users are selected, vetted and controlled and ensuring that participants have access to appropriate information, control over their data in the form of granular consent options and an ongoing right to opt-out. These factors were seen to have been key to the ongoing trust relationship between the focus group participants and ALSPAC.

Annex 3: Evidence on the collection of spatially indexed and location records

A3.1 Introduction

ALSPAC systematically records participant address details in order to administer study activities and as a means to conduct spatially informed research. The study has compiled spatially and temporally indexed information, including: area-level built and social characteristics; exposure measurements; participant-reported data directly related to the spaces and places they inhabit; and, information directly measured from participants (e.g. blood lead and total mercury measurements)[44]. These data collection strategies have involved seeking self-report information from participants or from using their location – typically residential address – to link to existing information, or to model new information. However, the ubiquity of personal computing and linked sensors introduces new possibilities for highly granular data collection at an individual rather than ecological level or residential address level. Examples of this could include personal physical exercise monitoring (e.g. linkage to 'Fitbit' or 'Strava' exercise app records), air pollution exposure (through personal air pollution sensors) and spatial and temporal movement through the environment (through GPS location monitoring). Yet this increasingly granular data collection may generate concerns regarding intrusion and disclosure risk, and there was limited information to understand whether this use of participant information is within the reasonable expectations of those enrolled in the study.

A3.2 Methods

ALSPAC data managers attended an OCAP meeting (2017) to seek views on the study's use of personal location data; whether this type of research is viewed as important and within the perceived scope of ALSPAC; and, whether participants were concerned about the use of their location information, or perceived specific risks to this type of research. To facilitate the discussion, the data managers presented hypothetical research scenarios that described sharing approximate location (e.g. 1 km2 area), specific locations (e.g. home or school addresses) and exact location (e.g. GPS tracking). OCAP members unable to attend were provided with the information and were able to provide written comments. Two OCAP members summarised the OCAP views, with the full group approving their final text.

A3.3 Results

The following results are summarised from Boyd et al. (2019).

OCAP viewed location-based research as broadly acceptable; but expressed preferences for the processing of the location data to be restricted to trusted study staff. Concerns were raised regarding GPS location tracing data, although there was recognition that many

[44] Boyd A, Thomas R, Hansell AL, Gulliver J, Hicks LM, Griggs R, Vande Hey J, Taylor CM, Morris T, Golding J, Doerner R, Fecht D, Henderson J, Lawlor DA, Timpson NJ, Macleod J. Data Resource Profile: The ALSPAC birth cohort as a platform to study the relationship of environment and health and social factors. *Int J Epidemiol*. 48:4, 2019. https://doi.org/10.1093/ije/dyz063

individuals were routinely tracked in this way by the commercial developers of device operating systems and applications. A more detailed summary of OCAP views has been developed by members of the committee and is re-published below (see Panel 1).

Panel 1: ALSPAC's use of participants' location data: a participant perspective*

Introduction:

ALSPAC data managers consulted the Original Cohort Advisory Panel (OCAP) aiming to understand: participant views on personal location data, whether this research is viewed as important and within the scope of the study; and if participants had concerns or perceived there to be risks to this type of research. Established in 2006, OCAP currently comprises around 30 participants (aged 25–27).

Methods:

In late 2017, data managers attended an OCAP meeting (members unable to attend were able to provide writ- ten comments). To encourage discussion, the data managers presented hypothetical research scenarios that described sharing approximate location (e.g. 1km^2 area), specific locations (e.g. home or school addresses) and exact location (e.g. GPS tracking). Two participants summarized OCAP views for this publication, with this text approved by the full group.

Results:

Regardless of the scenario presented, there was consensus that this type of research is important, particularly where the potential to improve public health was clear. Research using personal location data was perceived as different from other research, but within scope of the study. Several participants mentioned that the data that have already been collected should be made the most of. Many of the concerns raised could be addressed by standard safeguards that are in place for other types of ALSPAC data: for example, issuing contracts for data sharing, enforcing sanctions for misuse, and encryption of data. There was some discussion around the feedback of results to participants. Again, clarifying standard ALSPAC procedures resolved the questions; participants would not expect personal return of results and the benefits would be felt by wider society. A small number of participants expressed concerns about aspects of sharing approximate and specific location data. In general, the group were comfortable with the sharing of approximate location data and this was not perceived as being as personal as the other location data under discussion. However, a few participants remained concerned about the potential for identification where cell sizes were small. With regard to sharing specific location data, there was some indication that certain locations are perceived as more sensitive than others. For example, some participants expressed that they were more comfortable for their school address to be shared than their home address, owing to the number of other students at the school (though the question of small cell sizes arose again). There was some concern that the sharing of multiple locations would raise the risk of identification, and that conceptualizing certain locations as 'historical' is inappropriate as they may still be current for participants and their families. The biggest concern in relation to sharing specific location data was that multiple datasets could be linked through common variables, thus making identification more likely. Of course, this problem is not unique to ALSPAC, but also applies to many other longitudinal studies. Some participants felt reassured knowing that only bona fide researchers would be given access to these data. However, this issue remained a significant concern for a small number of participants.

Across the group there was less consensus with regards to collecting and sharing exact location-tracking (e.g. Global Positioning System) data. Some participants immediately found this acceptable whereas others did not. It was recognized that, as this would involve new data collection, participants could choose not to take part in this. One participant highlighted that new data collection would be scrutinized by an ethics committee and that their concerns lay more in the secondary access to these data. Some perceived harms were expressed by the group (such as the use of these data in legal cases). However, there was a general sense that many participants already face these risks in their day-to-day lives owing to commercial collection of location data. Indeed, it was suggested that participants might find this type of data collection more acceptable because of familiarity with this type of data collection. In general participants were not concerned by sharing events (e.g. that they passed a certain natural feature) but some had reservations about sharing the location (e.g. that they were on a particular road when they passed it). Some participants had particular concerns when it came to these data being connected to their children. Despite seeing the value in sharing these exact location data and perceiving it as within scope, there remained some concerns, and it was not always easy for participants to rationalize or articulate why the idea did not sit comfortably with them.

Conclusions:

Five key issues came to light during the overall discussions: (i) the suggestion of using a split processing approach (as described in the main article text) was generally well received and preferred across a majority of scenarios; (ii) separately, there seemed to be a general preference for steps in research that involve processing the personal location data to be done in-house at ALSPAC, though there was also recognition of the significant burden this would place; (iii) in general, participants wanted to know that this type of research is taking place; (iv) in a majority of research scenarios, some type of consent process was expected, with an opt-out campaign receiving generally positive views and being thought of as in keeping with previous campaigns in ALSPAC (e.g. for recall by genotype studies); v) the extent to which personal location data such as addresses are conceptualized as data rather than as a means for participation needs to be carefully addressed. Overall, there was consensus that the types of research enabled by use of personal location data would be important and within scope for the Study. A majority of participants seemed to agree that use of personal location data was acceptable given the safeguards that could be put in place, and that the benefits out- weigh the risks. Specific concerns differed between scenarios, suggesting that the safeguards that are put in place could vary in complexity on a case-by-case basis. The sharing of approximate and specific personal location data was arguably more acceptable than sharing exact location-tracking data. However, the discussion reveals that participants are at least willing to consider this option also. Underpinning the discussions was a sense of trust placed in ALSPAC by its participants.

* taken from Boyd et al. 2019. DOI: 10.1093/ije/dyz063. Reused under a Creative Commons BY-CC licence.

A3.4 Conclusions

Participants expect that confidentiality risks introduced by spatial research are controlled. Data processing should be designed using ‘split stage’ protocols to separate the processing of identifiable location data (e.g. lists of postcodes or grid co-ordinates) from the analysis process. Ideally, all processing of identifiable location data should be conducted by study data managers. However, where this is not possible – such as where particular expertise or equipment is required – then equivalent measures should be deployed. For example. the identifiers of study participants should be masked by purposefully selected equivalent identifiers of non-study participants. There is a need for transparency in research use of location-based research and that mechanisms need to exist for participants to opt-out. Further risk assessments and participant consultation is required before precise location tracking data collection is deployed: it is likely that the necessary safeguards needed to ensure acceptability will need co-designing with participants and experts in disclosure risk.

Annex 4: Evidence on linkage to Social Media records

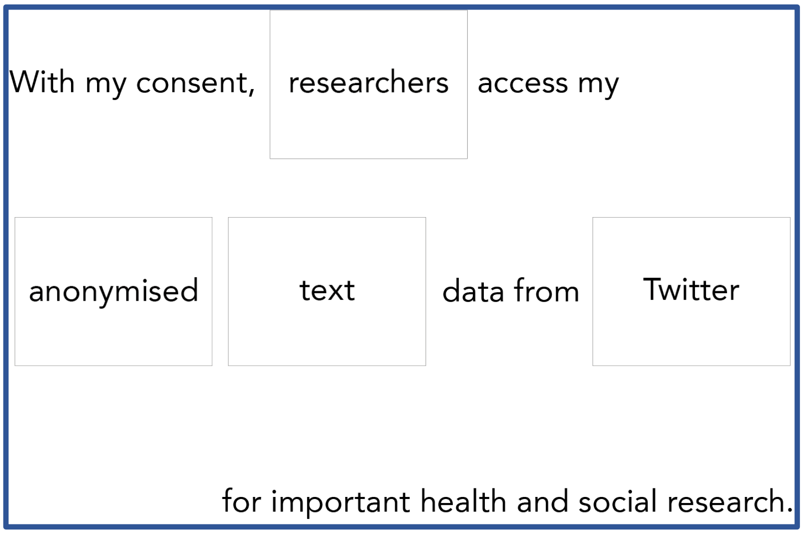
A4.1 Introduction

There is great potential for social media records to enhance cohort databanks through adding rich information about behaviour, emotions, aspirations and communications. There is also great potential for existing cohort study data to be used to provide ‘ground truth’ assessments of self-report and objectively recorded information that can be used to validate inferences made on social media data alone. Previous focus groups testing the acceptability of research using social media data have only assessed this use in general population samples. Participant’s with long-term engagement with cohort studies such as ALSPAC may have different views or levels of understanding.

A4.2 Methods

Two focus groups explored participant views towards the acceptability and necessary safeguards needed to support the use of social media data in research. ALSPAC participants over two generations - young people (N=9) aged 26-28 and parents (N=5) aged 53-65 - took part in two separate focus groups. The focus groups used semi-structured discussion to capture views. This was supported by the use of a phrase template as an elicitation tool. The template (see Figure 2) had some fixed text punctuated with blank spaces; separate word cards were provided and participants were asked to use the cards to fill in the blanks (up to 108 possible scenarios) and discuss how they would expect these data to be shared and presented. Participants discussed a range of social media platforms, but were informed that access would only be to their ‘visible’ information (i.e. what the public or their permitted friends and family could see), not their private information or other data stored on their systems (e.g. in online cloud storage associated with their social media).

Figure 2: An example of a completed template from the elicitation exercise.



With my consent, researchers access my

anonymised text data from Twitter

for important health and social research.

A4.3 Results

The following results are currently unpublished.

Participants were supportive of ALSPAC initiatives to collect social media data to facilitate health and social research; and felt their trust in ALSPAC enables this:

> *"'Children of the Nineties' is fine because I know you're not going to sell it" – Young Person*

> *"I would trust emphatically the 'Children of the Nineties'" – Parent*

The study parents also suggested that they self-censor their posts to only include information that they would not mind others seeing. However they expressed some concerns that sharing data with ALSPAC may expose them to IT security related risks, and these may relate to uncertainties or misconceptions about using software to conduct the extracts and how this would work in practice.

Participants discussed the parameters for the acceptable use of their social media data; with anonymity been seen as important:

> *"I suppose when things become anonymized it all seems a lot more fine. If it's being reduced to numbers and data points I would be much more likely to give my consent." – Young Person*

Concern was expressed about the use of private messaging data and third-party data (i.e. data from friends or connections who had not consented), this view was linked to etiquette and control:

> *"I object to it strongly... my friends haven't agreed to that." – Parent*

> *"I wouldn't be happy if someone consented on my behalf. And that's the same on every platform. It's not my place to consent on someone's behalf." – Young Person*

And that study extracted data may continue to be stored and used even where the original poster had deleted information on the social media channel:

> *"Twitter is public, but only for the time you leave it up. If you give permission for someone to store and access your data you're giving them permission to have it for as long as they want it." – Young Person*

Yet most considered the wide range of other data - including likes, network, text and location information – to be acceptable. There were mixed views on the use of photos, with the parents being willing for these to be shared on the basis that they only select photos they would be happy for others to see and share; but some young people expressing concern about the inclusion of third-parties. Location data was explicitly discussed but not seen to be concerning or any more sensitive than the other data (although some suggested they did not enable this feature in their social media accounts). These views did not seem to differ by social media platform type.

The word template exercise established that participants were accepting of either an automated program harvesting their raw social media posts or ALSPAC Data Managers conducting and overseeing this role. This acceptance did not extend to research users; here participants were accepting of anonymised use of processed data.

A4.4 Conclusions

Participants were accepting of ALSPAC's use of their social media records, except private channel information and information about third-parties. This acceptance was based on trust in the study, belief in the purpose for which the data would be used and reassurance about ALSPACs data management processes maintaining their confidentiality. Participant views did not tend to vary across the different social media platforms. There were some minor differences between the generations on their views of social media and how their social media data could be used for research. These differences may be influenced in the way different generations use social media, and by differing conceptions of how systems worked.

Annex 5: Evidence on an Ecological Momentary Assessment (EMA) of alcohol consumption.

A5.1 Introduction

Smartwatches enable us to retain the benefits of EMA, while reducing the disruption associated with smartphone based EMA data collection. As they are worn on the wrist, they are never beyond reach, so time taken to access the device is significantly reduced. In addition, because they are worn against the body, less intrusive haptic prompts can reliably be used. The development of the microinteraction-based EMA (μEMA) techniques utilised in this study goes further in reducing disruption, by reducing requests to a single question with a limited set of answers that can be responded to with a single tap. When compared with smartphone-based EMA over a 4-week period, smartwatch-based μEMA had better compliance (82% v 64%), completion (92% v 67%), and lower levels of disruption (38% v 53%).[45]

The proposal was for participants to be issued with an ALSPAC Smartwatch and for this to be used to collect self-reported EMA data on consumption of alcoholic drinks, and this was explained in depth to participants. The proposal was for ALSPAC to issue a smartwatch to participants which should be worn daily for three months. Participants would be sent notifications every two hours (past 12:00) asking if they had drunk an alcoholic drink in the last two hours. If the participant responded positively then they would be asked a range of follow-up questions: what type of drink [beer, wine, spirits, other], size of drink? [1/2 pint, pint, 330ml bottle], How many drinks have you had? Are you at home or outside the home? Are you with people or alone? The smartwatch EMA would be accompanied by online self-report questionnaires collecting data on the same topic. The study would be consented and would use a recall-by-phenotype case selection which would select within participants reporting a minimum level of alcohol consumption.

A5.2 Methods

The project was discussed at the OCAP meeting in January 2019. There were six OCAP members present at the meeting. The meeting was chaired by a member of the ALSPAC participation team and the study PI (author AS) was present at the meeting to outline the projects aims and methods. AS brought an example smartwatch to the meeting to illustrate the EMA system. This was a slightly older style smartwatch, and AS indicated the smartwatch used in the study would hopefully be an up to date model. AS described the project and this was followed by a discussion comprising questions from OCAP and answers from AS. Thematic analysis and conclusions were compiled based on the minutes of the meeting.

[45] *Intille, S., Haynes, C., Maniar, D., Ponnada, A., & Manjourides, J. (2016). μEMA: Microinteraction-based Ecological Momentary Assessment (EMA) Using a Smartwatch. Proceedings of the ACM International Conference on Ubiquitous Computin . UbiComp (Conference), 2016, 1124-1128.*

A5.3 Results

The following results are currently unpublished.

The questions and discussion subject matter broadly focused on the use of a smartwatch for an EMA of alcoholic drinking habits, although other use cases were briefly touched on. Relating to the alcohol use case, the discussion can be grouped into the following themes: questions about how the device/data collection would work (see Table 1); appearance and acceptability; Practicalities of completing the EMA; and, scientific issues.

Table 1: Questions and answers about the proposal which framed the debate.

Questions from OCAP	Answers from study PI
Would the data collection software be added to their own device or would they be provided with an ALSPAC device?	ALSPAC would provide participants with a bespoke smartwatch.
What operating system does it use?	Android.
How do you charge it?	Like an 'Apple' watch.
What if you forget to wear it?	The 'buzzing' reminder may remind you.
Are you going to ask people before the study how much they drink?	This will be a recall-based case selection using information from existing questionnaire data to only include those reporting a minimum amount of alcohol consumption.
When is it due to go live?	August 2019.
How do you get the data?	The watch is returned after three months and then the data is downloaded.

The substantive elements of the OCAP discussion focused on three core themes:

- **Appearance and acceptability:** members commented that smartwatches were used by lots of people and the example device "looked like an Apple watch" so would probably be seen as acceptable. They were asked if they would rather use their own personal device, the members suggested they would: although it was noted by the PI that the software was currently tied to the Android platform. There were questions about the reminder trigger, and if this were a loud alarm would this be disruptive? A participant noted the 'buzzing' of the vibration reminder was still quite loud.

  OCAP members considered EMA questionnaires to be appropriate and of an acceptable length and it was quick and simple to complete. They noted the 'buttons' on the survey were large and easy to use. They suggested the study was quite demanding on their time, particularly the requirement to undertake the EMA for three months and to complete self-report questionnaires: suggesting that only a pilot group should be required to complete both types of assessment. Further suggestions were that the questionnaire should be online and perhaps should only be required at the beginning and end of the three-month period.

OCAP asked about feedback of results and suggested that a linked app may be a desirable feature as participants might like to look at their data. It was suggested that 'well done' messages at intervals may help with adherence, although these should not be intrusive. It was suggested that there should be staged 'compensation' payments of £10 per month with an additional £10 for completing all three months.

OCAP asked about whether this could be considered 'spying' and if there were safeguards for alcoholics. Members seemed reassured that this was not 'spying' given it would be opt-in with informed consent using transparent participant information materials. There was not a fully developed plan for screening for alcoholics at this point, so this was not fully explored. However it was made clear we would be consulting expert advice on this matter.

- **Practicalities of completing the EMA:** OCAP members focused on the alcohol consumption EMA. After discussing the practicalities of the survey (e.g. when would you be prompted to provide data, how the question set was compiled) the OCAP members discussed practical design features and potential improvements. They stated the need for a 'back' button to be able to correct mistakes. This was linked to a suggestion that there could be an option at the first reminder to say you would not be drinking alcohol at all that day, with the back button providing the means to change this answer. There was strong feedback about the need for a 'repeat' option to provide a shortcut during a "big night out". It was also suggested there should be a 'skip' option to allow the survey to be completed a bit later on – or possibly the next day – if it wasn't convenient at the initial moment (e.g. the wearing was currently dancing). There were several comments on how different drinks would be classified, for example cocktails and the differences between drinking spirts as a 'shot' (typically 25mm volume) or as a 'long' drink (i.e. a shot mixed with a non-alcoholic drink, such as a Gin & Tonic or Rum & Coke). There needed to be clarity in the guidance where you are pouring your own spirits or wine. OCAP advised that the questionnaire should capture information about special events, which could lead to different drinking habits. These could include their or their friends birthdays, weddings, festivals, and religious, cultural or popular social events that might impact on alcohol consumption, e.g. 'Dry January' and Lent.

- **Scientific issues:** OCAP discussed different aspects about the validity of the data and the potential for error. They discussed 'binge' drinking and the ability to provide accurate answers, and questioned whether people would answer truthfully. The discussion compared the benefits of being able to edit the answers the next day to add missed information, but the dangers of recall error countered the benefits. OCAP questioned whether the act of recording alcohol consumption would lead to changes in behaviour. While the PI stated that evidence suggests this is not the case, this fed into comments about validation – including through biological assays (a fingerprick test was suggested).

A participant asked if a smartwatch could be used for sleep assessments, and on hearing that it could, they suggested that three months was too long a time period for wearing a device throughout the night. Another participant asked if the device had Global Positioning

Software capabilities (i.e. would it record location). The study PI responded that it did not have this capability and the issue was not explored any further.

A5.4 Conclusions

Overall, the OCAP members suggested they were broadly supportive of this proposal. There was limited discussion of risks, and this may be a result of this being a consented ALSPAC-centric exercise where the study controlled all aspects of the process, including the device and data. OCAP's guidance suggests the software is broadly acceptable with a strong recommendation that there needs to be a 'repeat' function to make completing the survey easier. The guidance that the members would rather use their own device suggests the need to develop software for all the major platforms. This was coupled with suggestions for automatic data uploads. Although, as the overall take-up of smartwatches amongst the general population is relatively low, it suggests an ongoing need for study provided devices.