Five models for child and adolescent data linkage in the UK: a review of existing and proposed methods

Karen Laura Mansfield ![ORCID],[1,2] John E Gallacher ![ORCID],[1] Miranda Mourby ![ORCID],[3] Mina Fazel ![ORCID] [1,4]

[1]Department of Psychiatry, University of Oxford, Oxford, UK
[2]Oxford Health NHS Foundation Trust, Oxford, UK
[3]Centre for Health, Law and Emerging Technologies (HeLEX), Faculty of Law, University of Oxford, Oxford, UK
[4]Children's Psychological Medicine, Oxford University Hospitals NHS Foundation Trust, Oxford, UK

Correspondence to
Dr Karen Laura Mansfield, Department of Psychiatry, University of Oxford, Oxford OX3 7JX, UK; karen.mansfield@psych.ox.ac.uk

Check for updates

ABSTRACT

Over the last decade dramatic advances have been made in both the technology and data available to better understand the multifactorial influences on child and adolescent health and development. This paper seeks to clarify methods that can be used to link information from health, education, social care and research datasets. Linking these different types of data can facilitate epidemiological research that investigates mental health from the population to the patient; enabling advanced analytics to better identify, conceptualise and address child and adolescent needs. The majority of adolescent mental health research is not able to maximise the full potential of data linkage, primarily due to four key challenges: confidentiality, sampling, matching and scalability. By presenting five existing and proposed models for linking adolescent data in relation to these challenges, this paper aims to facilitate the clinical benefits that will be derived from effective integration of available data in understanding, preventing and treating mental disorders.

INTRODUCTION

Adolescence is a critical period for the emergence of mental disorders,[1 2] and there is evidence to suggest that adolescents are presenting to mental health services in increasing numbers,[3] with rates increasing most for those between the ages of 15 and 18 years.[4–6] This can be explained in part by a greater acceptability of seeking care, but also by increases in prevalence of anxiety and mood disorders.[6] At present it is estimated that under 30% of those who need mental healthcare actually access it.[7] Studies have identified how adolescents who access support for mental disorders early have improved functional and academic outcomes at age 18 years,[8] highlighting the importance of early identification and intervention. Without appropriate support there is a risk of downward-spiralling trajectories, with negative impact on the health, social, occupational and learning outcomes of the young person.[9] The absence of support can lead to ramifications for the young person, their immediate family and broader society, including demands on primary and secondary care and unemployment.

Many factors have been studied exploring associations of risk for mental disorders in young people, as well as the effects of specific interventions. Rather than there being isolated risk factors, it is more likely that mental disorders reflect the accumulation of multiple risk factors[10] as the developing mind probably depends on a dynamic interaction between both risk and protective factors (figure 1). In the UK, information relating to many relevant factors already exists in national routinely-collected data sets, such as health, education and social care records, which could be linked to provide information on additional variables and outcomes.[11 12] In addition to the administrative data, selected information from large-scale research cohorts could be incorporated to address specific research questions.[13–15] If such a triangulation of data could be facilitated, then researchers and clinicians would have the potential to investigate additional outcomes and to control for many factors that have previously been difficult to take into account and insufficiently powered in analyses.[16] Despite these opportunities, only a small number of studies in the UK have managed to use large routinely-collected datasets to investigate multifactorial influences on the developing mind, particularly when linking education data.[17–19] This likely reflects some key challenges to data linkage. This review aims to facilitate the clinical benefits of data linkage by describing current and hypothesised models of data linkage in relation to key challenges that such work presents.

METHODS

We focused on linkage between large-scale mental health, education and research datasets (table 1), due to their size and representativeness for adolescent mental health. In particular, data collected from schools by the Department for Education National Pupil Database (NPD) presents an ideal sample frame because it comes as close as is currently possible to whole population adolescent census data, although with limitations.[17 20] The NPD itself includes limited information on social care provision, but future linkage to social care records would provide further pertinent information. Linking all pupils with records in the NPD to information on adolescents referred to secondary mental health services enables longitudinal research that can address mental health from the population to the patient, by mapping the development of mental disorders. We first describe four key challenges to linking these important data sets, before presenting the potential models and how they address the challenges.

Key challenges to linking adolescent mental health data

Preserving privacy and confidentiality
This first challenge reflects the confidential nature of the information gathered and the need to maintain the anonymity of the individuals who might not have consented to participating in research.
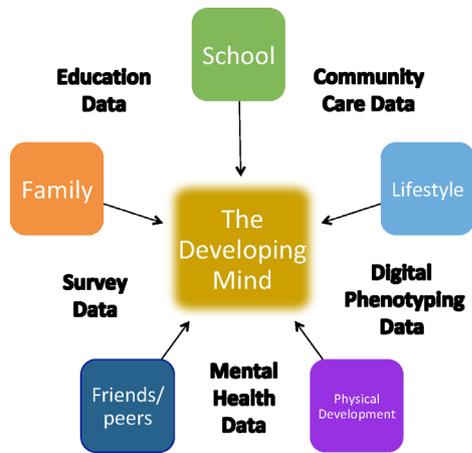
Figure 1 A graphical illustration of some of the major influences on the developing mind and some existing large-scale data sets that already capture related measures.

Although research can ultimately be performed on linked, anonymised data,[21] the prior process of linking separate records from health, education and research requires access to identifiers. This requires the processing of personal data, for which a 'lawful basis' under the General Data Protection Regulation must be identified and documented in advance.[22] Where significant amounts of special category data (such as identifiable health information) are processed, this must be done as part of a Data Protection Impact Assessment (General Data Protection Regulation (GDPR) Article 35). Although consent is conventionally sought for participation in research, the lawful basis for processing personal data for research is usually described as a 'task in the public interest' under UK guidance,[23] for which appropriate information governance (IG) and security controls need to be in place. Health data is a special case due to the highly sensitive nature of patients' medical history. Identifiable (NHS) health data in the UK may only be exchanged for research (without consent) with the approval of the Health Research Authority, on the advice of the Confidentiality Advisory Group

(CAG) to 'set aside the duty of confidence' (referring to s251 of the NHS Act 2006). There are solutions that attempt to avoid the processing of personal data altogether by linking 'de-identified' (pseudonymised) records, but under GDPR pseudonymised data are not generally considered to be 'anonymous' if they can still be linked back to individuals, and the IG advantages of these linkage solutions need to be considered in relation to which approvals are required.[24]

Acquiring a representative sample
Although the NPD itself contains a relatively complete population sample, which linkage method is used will impact on the representativeness of the sample.[25] For example, if the lawful basis for processing (linking) personal data is 'consent', then the linked sample will be less representative.[26] Being able to add research data is likely to be central to any successful linkage model, but this introduces similar limitations, especially in the case of those under 18 years of age. In most circumstances, collecting research data from children and adolescents requires explicit parental consent as well as active assent from the children and adolescents, which can then be further complicated if the study collects data beyond age 16 when the adolescent will need to give their consent anew. The consent procedure can reduce and bias the sample because of these practical limitations,[27] potentially compromising the validity of the findings. In certain research ethics procedures, parents can be provided information on the research and given the opportunity to 'opt-out', whereby consent is assumed. This method usually has considerably better recruitment than the 'opt-in' method.[28] However, there are strict guidelines concerning the anonymity of the data, making this route impractical for linkage to administrative datasets. There are exceptions for 'competent youths', whereby personal data may be processed with 'opt-in' consent from those aged under 18 years following 'opt-out' procedures with parents. Guidelines vary and are resource-intense as they are determined on an individual project basis as part of research ethics procedures.

Table 1 Large-scale digital data sets from health, education and research settings considered in relation to linkage in this review.

Type of data	Data set	Data controller(s)	Measures	Population sample	Measurement mode	Measurement frequency
Mental health data	Clinical Record Interactive Search (CRIS) data from NHS mental health trusts	Participating NHS trusts	De-identified data; support team, referrals, episodes, diagnoses, treatment, text notes	Children and adolescents using secondary care mental health support from Child and Adolescent Mental Health Services (CAMHS) teams	Electronic Health Records (EHR) for CAMHS	Per clinical contact
Community care (health) data	National Child Measurement Programme (NCMP)	NHS Digital	Height, weight, BMI	Children attending schools and in reception or year 6 at time of measurement (annual)	Community care teams or school health nurses	Reception and year 6
Education (school) data	National Pupil Database (NPD)	Department for Education (DfE)	Attainment, absence, exclusions, free school meals, children in care, SEN, etc.	Pupils aged 3–19 from all maintained primary, secondary and special schools	School administrative systems and databases	Annual (some fields 3x per year)
Survey (research) data	Pupil Survey on Health and Well-Being	Local authorities, NHS trusts, or universities	Validated and unvalidated mental health questionnaires, risk and protective factors, etc	Consented pupils from participating schools	Surveys in schools, usually online	Variable—from one-off to termly/annual
Digital phenotyping (research) data	Remote Digital Phenotyping	Universities, and/or NHS trusts	Digital measures related to mood, movement/actigraphy, heart rate, etc	Consenting participants/ patients (with parental consent)	Websites, phones and/ or wearable devices	Weekly, daily or continuous

BMI, Body Mass Index; ID, Identifier; NHS, National Health Service; SAIL, Secure Anonymised Information Linkage; SEN, Special Educational Needs.

Matching

Matching individuals between datasets presents technical challenges. NPD and NHS datasets are recorded in separate databases and, although they both use unique identifiers, the unique identifiers are different in each case—pupil numbers versus NHS numbers. Matching is further complicated by additional identifiers if integrating research and social care datasets. Matching individuals using identifiers that are not unique (eg, name, date of birth, postcode) limits the practicality, accuracy and number of matches.[17 29 30] There are theoretical means to reducing the chance of unmatched cases, particularly if the data could be linked at a national level, but solutions to this problem are likely to be complex and resource-intensive.

Scalability

If a linkage model could be defined that can be scaled to a national level, this would facilitate a secure and streamlined application and linkage process. Currently, data application procedures for access to administrative data sets can involve filling out multiple and long application forms, complicated often by additional processes and possible delay,[31 32] particularly when linking multiple data sets. For example, overcoming the obstacles to linking health data in particular will require collaboration between a number of organisations.[33] The complexities include determining which ethical approvals need to be sought, which organisations can grant those approvals, and, when multiple UK countries are involved, if these processes need to be done for each devolved nation. Although there are discussions to try and harmonise these application processes, the pace of change is slow.

PRESENTATION

The models described below have been identified through a process of literature research, and working with data controllers, researchers and stakeholders. The core components described relate to differences in whether matching the data requires personal data to be exchanged and the lawful basis for processing personal data. We outline the models, giving concrete examples where possible, and describe how they address the challenges of confidentiality, sampling, matching and scalability.

Model 1: exchanging personal health data with CAG approval

This model has been used for linking child and adolescent mental health service (CAMHS) data held in the Electronic Health Records (EHRs) of the South London and Maudsley NHS Foundation Trust (SLaM) to school attendance records (NPD). Downs *et al*[17] sought CAG approval to set aside the duty of confidence, in order to send identifiable health data from SLaM records to the Department for Education (DfE). DfE then looked for matches in the NPD records based on name, date of birth and postcode, first looking for exact matches and then using 'fuzzy' matching to attempt to match cases that had been missed due to data entry errors. They successfully matched 82.5% of adolescents registered in SLaM to NPD.

Challenges: Downs described how it took almost 4 years to address the ethical, governance and technical challenges to achieve the linkage via this method, including having their first application rejected by CAG.[17] In this model, the sample frame is representative providing access is granted to data from the full (NPD) population living in the same region (not only successfully matched individuals), although adding research data would introduce limitations associated with consent. The probability of matching individuals accurately across data sets is possibly higher when the data processor has access to all identifiers in both data sets. In terms of scalability, CAG approval to set aside the duty of confidence is determined on an individual project basis, and therefore, if not sufficiently resourced, could become unmanageable.

Model 2: matching de-identified data by a third party

This linkage method has been used by the SAIL databank[30 34 35] and more recently by the Adolescent Mental Health Data Platform (ADP),[36] both based at the University of Swansea in Wales. ADP combines a number of different datasets collected on children across Wales, and links them using an Anonymous Linking Field (ALF), a form of Privacy Preserving Record Linkage (PPRL). PPRL uses a hashing algorithm to calculate pseudonyms, based on a selection of identifiers that are present in all data sets. The pseudonyms are used to match records, but cannot be translated back to the identifiable information because they are securely encrypted and the key is held separately by NHS Wales.

Challenges: PPRL offers a means to matching records while maintaining confidentiality, but it is not used widely outside of Wales. To our knowledge, the PPRL model has not yet been used as an alternative to CAG approval for linking NHS data from England to education (NPD) data, although it has been used as a secure method for linking NHS data from England to research data (eg, Clinical Record Interactive Search (CRIS) Network data linkages, https://crisnetwork.co). The sample frame is theoretically the same as when processing identifiable data with CAG approval, but might be affected by technical challenges: without the data processor having full access to all non-unique identifiers, it is difficult to match cases that are not an exact match based on the pseudonyms provided, and not as easy for the data processor to check the accuracy of the matching. However, this model does present a scalable solution because it reduces some of the concerns around confidentiality and has already been used at scale in Wales.[18]

Model 3: exchanging personal data with (parental) consent

A third model to linking health and education data to research has been used by the Avon Longitudinal Study of Parents and Children (ALSPAC) study, a large longitudinal research cohort incorporating a broad range of health-related measures from parents and their children.[37] By seeking consent from young cohort members (and their parents) to link their research data to routinely-collected data, ALSPAC data have been linked to NPD and CRIS (http://www.bristol.ac.uk/alspac/researchers/our-data/linkage/). This method has also been used to enhance the Millennium Cohort Study with primary and secondary health-care data,[13] with ongoing linkage to the NPD.

Challenges: If the lawful basis for processing personal data is 'consent', then preserving confidentiality rests on information security. However, the sample will be limited in numbers when compared with the other linkage models, reflecting variable response rates associated with seeking consent for research,[38] and be prone to further reductions due to withdrawal, attrition and the need to re-consent when adolescents become 'adult' (currently 16 years in the UK for research). The technical challenges might be fewer than for the first two models if the study team acquires sufficient details, including previous addresses. It is important to note that when linkage itself is based on consent, there needs to be a well-defined process in place to exclude data from individuals who later withdraw their consent. Therefore, scaling the consent model purely for linkage can be costly in

both time and resources, but linking an existing cohort to health and education data adds valuable dimensions.

Model 4: matching personal data within a local authority

This is a conceptual model, but to our knowledge some local authorities (LAs) are pursuing the possibility of linking measures collected from surveys to the education data they hold. LAs often hold census data from locally-maintained schools in their county, and some of them collect (anonymous) survey data in schools, including mental health measures. Some LAs also work with CAMHS and use data collected by school health nurses (such as the National Child Measurement Programme (NCMP)) to improve services. Researchers could work with local authorities to help them collect and link measures to guide service developments and policy.

Challenges: If all measures are collected by the same data controller, then personal data need not be exchanged before the data are anonymised for research. The sample might be limited by the fact that when schools become more autonomous academies, they no longer need to provide their data to the LAs. This makes it difficult for the LA to involve those schools in research, to access their data or link measures. When collecting identifiable survey data, it is not clear whether an LA would require explicit consent from parents, but this is more likely when the data concern adolescents' health. Matching the data could be facilitated if pupils can be logged in to a survey securely using a National Pupil Number and at least one other identifier. This model is scalable because LAs already submit data to the DfE/NPD, making it a relatively straightforward task to submit additional data collections for linkage.

Model 5: matching personal data within an NHS Trust

This model draws upon opportunities stemming from expanded child health services. For example, Oxford Health NHS Foundation Trust provides traditional CAMHS, as well as in-school mental health workers and school health nurses, with most data held in the EHR. Integrating additional educational measures into the EHR would be valuable, not only for the individual patient, but also for clinical research. NHS trusts often hold other valuable data that could also be linked to form a more comprehensive picture, such as digital phenotyping for online self-management systems like True Colours.[39]

Challenges: Similar to model 4, personal data can be linked by a single data controller (the NHS trust). The sample might be limited to adolescents who access CAMHS, including via mental health support in schools, without important education information that is included in the NPD. However, both the sample and the measures could be broadened using linkage to community care data (eg, NCMP) and potentially other (survey) measures collected by school health nurses. The technical challenges associated with matching the data would be minimal if all independent data collections could include NHS numbers—although to our knowledge this is not always the case. This model is certainly scalable and at least twelve NHS mental health trusts already make de-identified data from EHRs available for research via secure remote access as part of the CRIS network.

DISCUSSION

This paper presents five models of large-scale, cross sector data linkage (figure 2), with consideration to four key challenges: confidentiality, sampling, matching and scalability, with the goal to facilitate the clinical benefits of data linkage and to inform continued development of linkage models and data platforms for
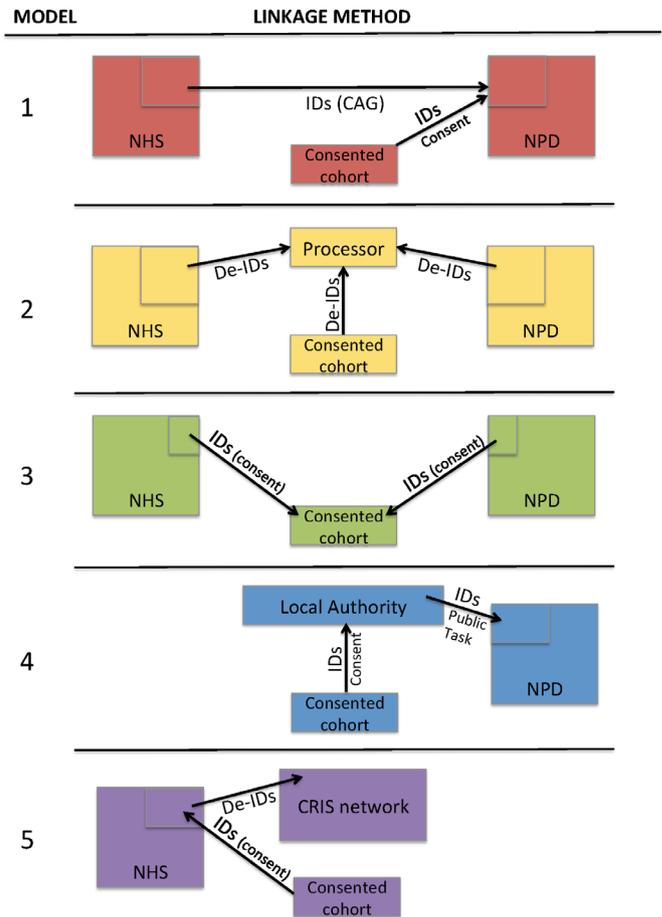

Figure 2 A simplified illustration of the linkage method for each linkage model. Arrows denote the direction of data-sharing for identifiable (IDs) vs. de-identified (De-IDs) data, for which the lawful basis for processing identifiable data (consent/CAG/public task) is shown. For simplicity, 'NHS' denotes NHS Digital or a NHS trust (data and data-sharing team), NPD denotes National Pupil Database (data and data-sharing team), and 'consented cohort' denotes research data (e.g. from surveys or remote devices). CAG, Confidentiality Advisory Group; CRIS, Clinical Record Interactive Search; NPD, National Pupil Database.

UK adolescent mental health research. Lessons can be learnt from each method. These include: capturing the NPD sample frame; identification of practical challenges; the advantages of the PPRL method to minimise the sharing of personal data; the richness of the measures when adding administrative data to a large-scale research cohort; and the unique opportunities of working with local authorities and NHS trusts who can potentially link health, education and research data within one environment. There are likely to be more potential models and challenges that have not been included in this review, but with the growing number of teams working on this issue across the UK, more insights into the practicalities and challenges of data linkage will become apparent and help provide further solutions.[32 33 40]

A successful model might depend on whom adolescents (and their parents) are most likely to trust. For example, individuals might be more willing to have their data linked by an independent data processor, so that the linked data will be anonymous to all parties. In this case a PPRL method for linking the data would be suitable, in which end users should not be able to identify individuals in the linked data. On the other hand, adolescents might feel protected if trusted professionals have the potential

to identify those considered to be at 'severe risk' and offer them support, as has been suggested during a discussion with a Young People's Advisory Group. A related question is the extent to which adolescents will give honest answers to sensitive questions in surveys if their responses are not anonymous. This might depend on who is administering the survey (the LA, CAMHS, universities), similar to the finding that a consent decision can depend on who is asking.[41] Research investigating attitudes to data linkage for research,[42] and further work with stakeholders, could better guide these decisions.[40 43 44]

When making administrative data accessible for research, preserving confidentiality with adolescents is crucial. The guidelines from the Information Commissioners Office are to anonymise research data as early as possible,[45] but there remains discussion around when de-identified data can be considered 'anonymous',[24] particularly since the passing of the GDPR. Some data scientists have demonstrated that re-identification is highly probable in large datasets and suggest further technical solutions.[46] Rather than relying too heavily on de-identification, data protection must rely on a balance of information security and IG safeguards.

It is important to consider how to maximise the value of the linked data for research and clinical care. There are future challenges that have not been discussed here, such as the limitations of categorical measures in administrative data and data harmonisation. Additional lessons can be learnt from the informatics architecture of established data platforms like the Dementias Platform UK,[47] which includes technical solutions for data protection (eg, split file double encryption), secure access, data curation, interoperability and analytical tools. In considering where to host the linked data, both trust and practicality need to be taken into account, particularly in scaling up the linkage to a national level. For example, NPD data can be accessed in anonymised form via the Office for National Statistics (ONS) Secure Research System.[48] ONS are also permitted to process identifiable health data, which could make model 1 a scalable solution without the need to seek approval from a CAG. However, this implies that ONS would theoretically hold identifiable linked data (even though access for research is in anonymised form), although it is yet to be determined whether this would be an acceptable option.

Although some cohort studies have aimed to include multi-modal data from the outset, data linkage offers a means to creating such cohorts at a large scale. With such rich data and sufficient power, the opportunities to better understand risk and protective factors become endless. For example, sophisticated algorithms, such as those that rely on neural networks and deep learning, could be developed to take all relevant measures into account.[49] Artificial Intelligence can be used to identify which combinations of the available measures can best be used to calculate 'risk' scores, to measure the impact of interventions, and even to predict outcomes for multiple specific combinations of risk and treatment, which could in turn inform more precisely targeted treatment. The linked data can also be used to track the efficacy of treatment and services. Analyses can be performed to assess critical ages and symptoms associated with mental health crises, and to develop effective screening measures of mental health that could be incorporated into service providers' own data via surveys. Future work will identify the next set of challenges and expand on these potential models to include other relevant data, such as social care and primary care records.

The clinical impact of the slowly-growing number of research platforms in the UK holding linked data relevant to adolescent mental health will benefit in particular from improved guidelines around the extent to which pseudonymised, linked administrative data can be classed as 'anonymous', and from research investigating which organisations (young) people trust to hold these linked data, in either identifiable or non-identifiable form. Further impact will be seen on future generations, when schools, local authorities, NHS trusts and mental health professionals are able to use algorithms and measures developed by others on linked data to maximise the value of their own data. This could help to detect risk factors, tailor services, prevent serious mental disorders and eventually reduce service utilisation as well as avoidable suffering.

Twitter John E Gallacher @dementiasUK and Mina Fazel @minafazeloxford

ORCID iDs
Karen Laura Mansfield http://orcid.org/0000-0003-0342-7926
John E Gallacher http://orcid.org/0000-0002-2394-5299
Miranda Mourby http://orcid.org/0000-0001-9323-9413
Mina Fazel http://orcid.org/0000-0001-9342-2365

REFERENCES

1 Gore FM, Bloem PJN, Patton GC, *et al*. Global burden of disease in young people aged 10–24 years: a systematic analysis. *Lancet* 2011;377:2093–102.
2 Patton GC, Sawyer SM, Santelli JS, *et al*. Our future: a Lancet commission on adolescent health and wellbeing. *Lancet* 2016;387:2423–78.
3 Fazel M, Rocks S, Glogowska M, *et al*. Reorganisation and investment in child and adolescent mental health services to improve access: a four-year observational mixed-methods study in England. Under Review.
4 Sadler K, Vizard T, Ford T, *et al*. *Mental health of children and young people in England, 2017. Trends and characteristics*. Government Statistics Office: NHS Digital, 2018.
5 Pitchforth J, Fahy K, Ford T, *et al*. Mental health and well-being trends among children and young people in the UK, 1995–2014: analysis of repeated cross-sectional National health surveys. *Psychol Med* 2019;49:1275–85.
6 Collishaw S. Annual research review: secular trends in child and adolescent mental health. *J Child Psychol Psychiatry* 2015;56:370–93.
7 Department of Health and Social Care, and Department for Education. *Government Response to the Consultation on Transforming Children and Young People's Mental Health Provision: a Green Paper and Next Steps*. London: Department of Health, 2018.
8 Neufeld SAS, Dunn VJ, Jones PB, *et al*. Reduction in adolescent depression after contact with mental health services: a longitudinal cohort study in the UK. *Lancet Psychiatry* 2017;4:120–7.

9 López-López JA, Kwong ASF, Washbrook E, *et al*. Trajectories of depressive symptoms and adult educational and employment outcomes. *BJPsych Open* 2020;6:E6.

10 Shulman C. *Research and practice in infant and early childhood mental health (children's well-being: indicators and research)*. Springer, 2016.

11 Einav L, Levin J. The data revolution and economic analysis. *Innovation Policy Econ* 2014;14:1–24.

12 Wijlaars LPMM, Gilbert R, Hardelid P. Chronic conditions in children and young people: learning from administrative data. *Arch Dis Child* 2016;101:881–5.

13 Tingay KS, Bandyopadhyay A, Griffiths L, *et al*. Record linkage to enhance consented cohort and routinely collected health data from a UK birth cohort. *Int J Pop Data Sci* 2019;4:10.

14 Holman CD, Bass JA, Rosman DL, *et al*. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Aust Health Rev* 2008;32:766–77.

15 Hsieh C-Y, Su C-C, Shao S-C, *et al*. Taiwan's National Health Insurance Research Database: past and future. *Clin Epi* 2019;11:349–58.

16 Hagger-Johnson G. Opportunities for longitudinal data linkage in Scotland. *Scott Med J* 2016;61:136–45.

17 Downs JM, Ford T, Stewart R, *et al*. An approach to linking education, social care and electronic health records for children and young people in South London: a linkage study of child and adolescent mental health service data. *BMJ Open* 2019;9:e024355.

18 Rahman MA, Todd C, John A, *et al*. School achievement as a predictor of depression and self-harm in adolescence: linked education and health record study. *Br J Psychiatry* 2018;212:215–21.

19 Booth JN, Leary SD, Joinson C, *et al*. Associations between objectively measured physical activity and academic attainment in adolescents from a UK cohort. *Br J Sports Med* 2014;48:265–70.

20 Jay MA, Mc Grath-Lone L, Gilbert R. Data resource: the National Pupil Database (NPD). *Int J Pop Data Sci* 2019;4:08.

21 Graham C. *Anonymisation: managing data protection risk code of practice*. Cheshire: Information Commissioner's Office, 2012. https://ico.org.uk/media/1061/anonymisation-code.pdf

22 Information Commissioners Office. Lawful basis for processing. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/ [Accessed 16 December 2019].

23 UKRI. GDPR and research – an overview for researchers. Available: https://www.ukri.org/files/about/policy/ukri-gdpr-faqs-pdf/ [Accessed 8 Jan 2020].

24 Mourby M, Mackey E, Elliot M, *et al*. Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK. *Computer Law Security Rev* 2018;34:222–33.

25 Downs J, Setakis E, Mostafa T, *et al*. Linking strategies and biases when matching cohorts to the National Pupil Database. *Int J Pop Data Science* 2017;1:348.

26 Mostafa T. Variation within households in consent to link survey data to administrative records: evidence from the UK millennium cohort study. *Int J Soc Res Methodol* 2016;19:355–75.

27 Shaw T, Cross D, Thomas LT, *et al*. Bias in student survey findings from active parental consent procedures. *Br Educ Res J* 2015;41:229–43.

28 Berry JG, Ryan P, Duszynski KM, *et al*. Parent perspectives on consent for the linkage of data to evaluate vaccine safety: a randomised trial of opt-in and opt-out consent. *Clin Trials* 2013;10:483–94.

29 Méray N, Reitsma JB, Ravelli ACJ, *et al*. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. *J Clin Epidemiol* 2007;60:883.e1–883.e11.

30 Lyons RA, Jones KH, John G, *et al*. The Sail databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 2009;9:3.

31 The Lancet Psychiatry. Smorgasbord or Smaug's hoard? *Lancet Psychiatry* 2019;6:631.

32 Iveson M, Deary I. Navigating the landscape of non-health administrative data in Scotland: A researcher?s narrative [version 2; peer review: 2 approved]. *Wellcome Open Res* 2019;4.

33 Mourby MJ, Doidge J, Jones KH, *et al*. Health data linkage for public interest research in the UK: key obstacles and solutions. *Int J Pop Data Sci* 2019;4:09.

34 Ford DV, Jones KH, Verplancke J-P, *et al*. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009;9:157.

35 Jones KH, Ford DV, Thompson S, *et al*. A profile of the Sail Databank on the UK secure research platform. *Int J Pop Data Sci* 2019;4:03.

36 The Platform. The adolescent mental health data platform. Available: https://adolescentmentalhealth.uk/Platform [Accessed 8 Jan 2020].

37 Golding J, Pembrey M, Jones R, *et al*. ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatr Perinat Epidemiol* 2001;15:74–87.

38 Boyd A, Tilling K, Cornish R, *et al*. Professionally designed information materials and telephone reminders improved consent response rates: evidence from an RCT nested within a cohort study. *J Clin Epidemiol* 2015;68:877–87.

39 Goodday SM, Atkinson L, Goodwin G, *et al*. The true colours remote symptom monitoring system: a decade of evolution. *J Med Internet Res* 2020;22:e15188.

40 Ford E, Boyd A, Bowles JKF, *et al*. Our data, our Society, our health: a vision for inclusive and transparent health data science in the United Kingdom and beyond. *Learning Health Systems* 2019;3:e10191.

41 Walker S, Potts J, Martos L, *et al*. Consent to discuss participation in research: a pilot study. *Evid Based Ment Health* 2019. doi:10.1136/ebmental-2019-300116. [Epub ahead of print: 26 Sep 2019].

42 O'Brien EC, Rodriguez AM, Kum H-C, *et al*. Patient perspectives on the linkage of health data for research: insights from an online patient community questionnaire. *Int J Med Inform* 2019;127:9–17.

43 Jewell A, Pritchard M, Barrett K, *et al*. The Maudsley biomedical research centre (BRC) data linkage service user and carer Advisory group: creating and sustaining a successful patient and public involvement group to guide research in a complex area. *Res Involv Engagem* 2019;5:20.

44 Perez Vallejos E, Koene A, Carter CJ, *et al*. Accessing online data for youth mental health research: meeting the ethical challenges. *Philos Technol* 2019;32:87–110.

45 Elliot M, Mackey E, O'Hara K, *et al*. The anonymisation decision-making framework: UKAN Manchester, 2016. Available: https://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf [Accessed 8 Jan 2020].

46 Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* 2019;10:1–9.

47 Bauermeister S, Orton C, Thompson S, *et al*. Data resource profile: the dementias platform UK (DPUK) data portal. *BioRxiv* 2019;582155.

48 Department for Education. How to apply for data extracts from the National pupil database (NPD), school workforce, individualised learner record and higher education statistics agency, 2018. Available: https://www.gov.uk/guidance/how-to-access-department-for-education-dfe-data-extracts [Accessed 8 Jan 2020].

49 Rajkomar A, Oren E, Chen K, *et al*. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018;1:18.