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

Karen Laura Mansfield, John E Gallacher, Miranda Mourby, Mina Fazel

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, and there is evidence to suggest that adolescents are presenting to mental health services in increasing numbers, with rates increasing most for those between the ages of 15 and 18 years. This can be explained in part by a greater acceptability of seeking care, but also by increases in prevalence of anxiety and mood disorders. At present it is estimated that under 30% of those who need mental healthcare actually access it. Studies have identified how adolescents who access support for mental disorders early have improved functional and academic outcomes at age 18 years, 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. 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 as the developing mind probably depends on a dynamic interaction between both risk and protective factors. 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. In addition to the administrative data, selected information from large-scale research cohorts could be incorporated to address specific research questions. 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. 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 the whole population adolescent census data, although with limitations. 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.
Clinical review: Child and adolescent mental health

Although research can ultimately be performed on linked, anonymised data, 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. 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 ‘task in the public interest’ under UK guidance, for which processing personal data for research is usually described as ‘task in the public interest’ under UK guidance, for which

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. For example, if the lawful basis for processing (linking) personal data is ‘consent’, then the linked sample will be less representative. 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, 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. 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.

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

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

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.
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 healthcare 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.

**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 (e.g., 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 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.

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

<table>
<thead>
<tr>
<th>MODEL</th>
<th>LINKAGE METHOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NHS → de-IDs → Processor → NPD</td>
</tr>
<tr>
<td>2</td>
<td>NHS → De-IDs → Processor → De-IDs → NPD</td>
</tr>
<tr>
<td>3</td>
<td>NHS → IDs (consent) → NPD</td>
</tr>
<tr>
<td>4</td>
<td>Local Authority → NPD</td>
</tr>
<tr>
<td>5</td>
<td>NHS → CRIS network</td>
</tr>
</tbody>
</table>

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.
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.  

Research investigating attitudes to data linkage for research, and further work with stakeholders, could better guide these decisions.  

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, but there remains discussion around when de-identified data can be considered ‘anonymous’, 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.  

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, 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, NDPD data can be accessed in anonymised form via the Office for National Statistics (ONS) Secure Research System. ONS are also permitted to process identifiable health data, which could make model 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 multimodal 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.  

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 of 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  

Acknowledgements The authors would like to thank Taj Sallamuddin and David Newton for helpful conversations on the information governance and security considerations associated with data linkage.  

Contributors KLM, MF and JEG conceived the study. KLM drafted the manuscript. MF and JEG helped refine the manuscript and MM provided additional input on ethical, legal and governance aspects. All authors have read and approved the final version.  

Funding This paper represents independent research funded by an MRC Mental Health Data Pathfinder award to the University of Oxford (MC_PC_17215) and by the NIHR Oxford Health Biomedical Research Centre (BRC-1215-20005). MF was funded by the National Institute for Health Research (NIHR) Applied Research Collaboration (ARC) Oxford and Thames Valley. DPUI provided infrastructure for this project through MRC grant ref MR/L023784/2. The views expressed are those of the authors and not necessarily those of the MRC, NHS, the NIHR or the Department of Health and Social Care.  

Competing interests MF has an honorary contract with Oxford Health NHS Foundation Trust (OHHNSFT). KM is supported by the Oxford Health BRC, a collaboration between the University of Oxford and OHHNSFT.  

Patient consent for publication Not required.  

Provenance and peer review Not commissioned; externally peer reviewed.  

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study. There are no data sets associated with this manuscript.  

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.  

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/0000-0001-9342-1365  

REFERENCES  
Clinical review: Child and adolescent mental health