Towards early identification of mental health problems in children's social care

Lead organisation	Department of Psychiatry, University of Cambridge
Principal Investigator	Dr Anna Moore
Protocol Author(s)	Dr Anna Moore; Katherine Parkin; Dr Efthalia Massou
Study Design	1) Secondary analysis of longitudinal linked administrative data. 2) Risk prediction of child and adolescent mental health problems using machine learning models.

Contents of Protocol

Executive Summary

The purpose of this study is to advance the development of early identification tools for young people's mental health (MH) problems in social care settings. There are significant problems in identifying young people with MH problems in social care settings. This is due to a variety of reasons, including poor integration of administrative data held about young people. In particular, important indicators for MH problems may be noted in different electronic record systems (e.g. social care, education, and healthcare), and the significance of these risk factors (RFs) may only become apparent when these disparate systems are linked together and factors are viewed cumulatively. Other contributing issues are that young people are often not in contact with professionals who are able to accurately diagnose MH problems; for example, young people's points of contact may be teachers or GPs, rather than MH services [1]. Furthermore, access to Child and Adolescent Mental Health Services (CAMHS) for assessment, diagnosis and intervention is limited, and the professionals in other settings who interact with young people do not always know where best to signpost for appropriate support and intervention. These interrelated issues leave young people and families struggling without appropriate MH support [2, 3], and professionals such as social workers left with the difficulty of not being able to offer the support they would like to.

Without early identification methods, treatment delays occur, meaning that young people access support when their difficulties have worsened, which can hinder recovery. Early identification and intervention offer a more promising solution than later treatment [4]. Effective early identification tools which take into account MH-related RFs could give social workers more confidence in whether a young people has a MH problem, and offer signposting recommendations matched to risk level and types of problem, as well as highlighting which particular RFs are contributing to a young person's individual assessment/calculation of risk, allowing for personalised care and support approaches which target modifiable RFs.

We may address these problems by linking administrative data from social care, healthcare and education, allowing us to develop and test early identification tools for young people's MH problems. Through this approach, we can also measure the unrecognised MH needs currently being managed in social care settings, a figure which to date is difficult to ascertain. Furthermore, by using linked administrative data, we can map the prevalence and distribution of young people's MH problems and associated risk factors (RFs) in a region; this will allow the actual needs of the population to be compared to local service provision, and service gaps to be identified.

Our ultimate aim is to build a linked database known as Cam-CHILD (Cambridge Child Health Informatics and Linked Data), which will facilitate development of risk prediction models and early identification tools for young people's MH problems (known as 'Timely' tools). The Cam-CHILD database will integrate administrative data for the majority of young people aged 0 to 25 years in the Cambridgeshire and Peterborough region, with data coming from social care, healthcare and education, then being combined with a research bioresource. The linked database will exist in an identifiable format for clinicians including social workers to use when providing services for young people and families; a de-identified, near-real-time version will also exist for trusted researchers to use to address issues pertaining to care, support and health of young people. If shown to be effective and valid, early identification tools will be made available for use in social care and healthcare settings within Cambridgeshire and Peterborough, as well as nationally.

We will carry out preliminary analysis in a comparable database whilst Cam-CHILD is being built, with the aim of externally validating our findings and expediting analysis in Cam-CHILD. For this, we will use the ADP/SAIL database, a repository of anonymised, administrative data pertaining to social care, healthcare, and education, in order to explore the topic of young people's MH problems (ADP: Adolescent Mental Health Data Platform; SAIL Databank: Secure Anonymised Information Linkage Databank). Specifically, this database will be used to measure the proportion of unidentified cases of MH problems in social care settings. In addition, we will measure the prevalence and distribution of MH problems and >100 associated RFs in young people aged 0 to 17 years in social care settings and across Wales, UK. We will also explore relationships between MH problems and RFs. Subsequently, the ADP/SAIL database will be used to test if it is possible to develop algorithms and prototype early identification tools that enable us to accurately predict which young people develop MH problems in social care settings.

Aims of this study are:

1. To expedite the development of (i) Cam-CHILD, a linked database for measuring MH need and associated RFs across a multi-agency pathway, with a particular emphasis on social care and (ii) Timely tools, digital MH early identification tools for use in social care settings.

2. To use ADP/SAIL, a comparable database to achieve the above aim; database will be used to:
    - measure unrecognised MH need being met by social care;
    - develop methods to measure the levels of MH problems and associated RFs in social care settings, and across different regions of Wales, UK;
    - describe relationships between exposure to RFs and MH outcomes;
    - explore best methods for MH risk prediction and early identification algorithms;
    - apply these to the Cam-CHILD database, testing generalisability.

We will carry out the following steps:

1. Link ADP/SAIL datasets. Explore the quality of linkages and compare records of successful and unsuccessful linkages to understand any bias introduced through failed linkages.
2. Operationalise the measurement of RFs in each of the datasets, establishing proxy measures as required.
3. Map the prevalence and distribution of MH problems and RFs known to be associated with the development of MH problems. Explore how these distributions vary across Wales. RFs to be tested include specific demographic information, adverse childhood experiences, safeguarding alerts, patterns of service use, school attendance and attainment, among others.
4. By linking social care and healthcare datasets, explore the levels of MH need in social care settings.
5. Explore the relationships between the presence of RFs and developing a MH problem.
6. Develop prototype early identification tools by developing and testing risk prediction models.

Impact of the research: Data on the levels of MH need in young people who are in contact with social care is poor, and current figures are likely to be vast underestimates. Linking administrative data from social care, healthcare and education will allow us to provide a more accurate estimation of this unrecognised MH need in social care settings. Prediction models for MH problems will allow MH problems to be identified early, and proportionate support to be offered in a timely manner, facilitating greater recovery outcomes and promoting young people's and families' wellbeing. Moreover, accurate identification tools will facilitate smoother care pathways between social care and healthcare, and offer a tool to be utilised alongside clinical judgement to aid care and support planning with families. Mapping the prevalence and distribution of MH problems and associated RFs will provide valuable information to commissioners to inform service planning and resource allocation based on the needs of the local population.

Acronyms

ADP/SAIL	Adolescent Mental Health Data Platform [part of the] Secure Anonymised Information Linkage Databank (i.e. database relating to Welsh population)
AI	artificial intelligence
ALF	anonymous linking field (i.e. project-specific unique number assigned to people in SAIL databases to facilitate linkage)
Cam-CHILD	Cambridge Child Health Informatics and Linked Data (i.e. database being built in Cambridgeshire and Peterborough)
CAMHS	Child and Adolescent Mental Health Services (i.e. NHS services which assess and treat young people with mental health problems)
MH	mental health
ML	machine learning
NEET	not in education, employment or training
RFs	risk factors
VDI	Virtual Desktop Infrastructure (i.e. secure way of accessing SAIL data)

Background and Problem Statement

There are high levels of mental health (MH) need in children's social care settings [2, 3, 5, 6]. However, the data to estimate the actual level of need is very poor and existing figures are likely to be vast underestimates. In particular, poor integration of information held about young people makes it difficult to accurately estimate MH need in this population. Access to childhood MH support can be challenging, and there are even more barriers to access for young people within children's social care settings [7]. It is important to provide suitable MH support in a timely manner to those who need it. However, the current system is not set up to do this well, because it is unclear which interventions are most useful and there is no clear way to effectively identify young people who have MH needs in social care settings. Moreover, young people in these settings have distinct MH needs [8, 9] and there is some evidence that standard MH interventions may be harmful for some young people with a history of social care contact, for example, looked-after children [10]. As such, it is important to provide risk factor-informed interventions to this population (for example, specific trauma-informed and non-stigmatising interventions). Without this approach at present, outcomes for young people with MH problems in social care settings are poor, including high levels of deliberate self-harm, crises, behavioural difficulties, difficulties accessing education, long-term placements and NEET, all of which can lead to poor long-term health and social outcomes [11]. In summary, without an effective means of early identification, young people and their families can suffer for prolonged periods without suitable MH support [2, 3]. Furthermore, a failure to identify risk factors (RFs) and MH-associated problems early can delay treatment and lead to limited interventions failing to address significant causes of a young person's difficulties [4, 12].

In order to solve these problems, we need to: (i) accurately understand the incidence and distribution of MH problems and associated RFs in social care settings across different geographical regions; (ii) understand the specific relationships between RFs and MH outcomes; (iii) provide this information to commissioners so that they can match service funding to the specific needs of the local populations, and make evidence-based and targeted commissioning decisions, which offer a more effective use of the limited funds and resources (including staff) available for service provision; and (iv) develop reliable early identification tools, which do not rely on an already over-stretched CAMHS.

This will lead to: (i) quicker access to assessment and intervention; (ii) enablement of research into interventions for young people with MH problems in social care; (iii) clarity for social workers about young people who are challenging to diagnose and signpost; and (iv) facilitation of conversations about access to MH services. In turn, this will improve outcomes and experiences for young people and their families through better integration and access to MH services.

Aims and Objectives

The long-term research aims are to: (i) use linked administrative data from social care, healthcare, and education to develop methods of measuring the true impact of MH problems and associated RFs, in particular the measurement of unmet MH need within social care settings; (ii) build and implement validated algorithms for early identification of young people with MH problems, with particular emphasis on social care settings; (iii) identify early interventions that target underlying RFs; and (iv) develop support plans where early identification enables rapid RF-focused interventions.

Within the lifetime of this grant, the aims are to use ADP/SAIL to:

1. Expedite the build of a linked administrative database in Cambridgeshire and Peterborough by using ADP/SAIL database to refine methods to:
    a. Operationalise the measurement of MH and RFs within multi-agency data;
    b. Develop methods to map the prevalence and distribution of MH problems and associated RFs in multi-agency data;
    c. Estimate unidentified MH need within social care.
2. Explore relationships between exposure to RFs and MH outcomes.
3. Explore the best methods for developing accurate and usable child and adolescent MH risk prediction algorithms.
4. Begin applying these to the Cam-CHILD database, in order to test validity of the database and generalisability of the risk prediction algorithms.

The project objectives are to use ADP/SAIL to:

Objective 1: identify how MH problems and associated RFs are operationalised, in order to facilitate extraction and measurement within multi-agency data.

Objective 2: measure the prevalence and distribution of young people's MH problems across settings (social care, healthcare and education) and geographically-bound populations (across Wales, UK).

Objective 3: measure the prevalence and distribution of RFs associated with MH problems across settings (social care, healthcare and education) and geographically-bound populations (across Wales, UK).

Objective 4: measure the unidentified burden of MH problems within children's social care to explore how many young people with MH problems are being supported by social care.

Objective 5: describe the relationships between the presence of RFs and development of a MH problem.

Objective 6: explore whether it is possible to build accurate early identification risk prediction tools for use in social care settings, such that MH support can be targeted in an effective and timely manner.

Objective 7: apply this to the Cam-CHILD database, starting to test external validity and generalisability of algorithms for use in social care.

Research questions:

1. What is the best method of measuring MH problems and RFs for young people's MH problems in linked administrative datasets?
2. What is the prevalence and distribution of MH-associated problems and their RFs? How do patterns of MH-associated problems vary between social care, health, and educational settings? How do they vary across Wales, UK?
3. What is the unrecognised MH need in social care settings?
4. What are the relationships between RFs and MH problems?
5. What are the best methods for building predictive risk models and early identification tools for young people's MH problems for use in social care settings?
6. Can findings and methods be replicated across databases?

Method and Analysis

Study Design	A retrospective cohort study of young people aged 0-17 years will be used (objective 1). Measures of this cohort study will be used as a cross-sectional study to explore and describe objectives 2 to 5. Methodological approaches (i.e. splitting data into training and test set, testing recall of cases) using the same cohort study and particular elements of this (i.e. site-level data) will be used for objectives 6 and 7 correspondingly.
Population	Cohort definition:- Our population of interest is young people aged 0-17 years who are residing in Wales, UK. We will construct our cohort to mirror the cohort in Cam-CHILD by including records from any young people in Wales who were aged 0 to 17 years in the period between 1st January 2013 and 31st March 2020, and retrospectively including all data about those individuals

(e.g. if a young person was aged 12 on 2nd January 2013, they would be captured within our cohort and their data prior to this date would be included for analysis).

Representativeness of SAIL:-
The SAIL databank offers near-whole population-level data, with records for the majority of residents in Wales, UK (approximately 3 million people). In 2009, SAIL databank already had over 500 million anonymised and encrypted individual-level records from a range of sources relevant to health and well-being, and has continued expanding.[13] We will link data from the Welsh Demographic Service, Annual District Birth Extract, Annual District Death Extract and ONS Census 2011 in order to estimate the total Welsh population as accurately as possible, and will compare the other linked datasets to this to understand how representative our sample is of the Welsh population of young people.

Previous research in SAIL databank found that “The validation of using the NHS number yielded specificity values > 99.8% and sensitivity values > 94.6% using probabilistic record linkage (PRL) at the 50% threshold, and error rates were < 0.2%. A range of techniques for matching datasets to the NHSAR were applied and the optimum technique resulted in sensitivity values of: 99.9% for a GP dataset from primary care, 99.3% for a PEDW dataset from secondary care and 95.2% for the PARIS database from social care.” [13] As we are aiming to link more datasets, this may vary to an extent and we will explore whether there are statistically significant differences between the individuals whose records were successfully linked compared to those who were not, to better understand the representativeness of the created cohort. Based on previous research, we envisage that unsuccessful linkages will tend to occur more often for those who are from more disadvantaged backgrounds or who have moved around more as they may lack a unique identifier for deterministic linkage and may have insufficient identifiers to allow probabilistic linkage [14-16]. In order to mitigate this, we will set a relatively low linkage threshold which balances the risk of falsely linked records with the need to successfully link records from potential under-served communities.

Data

The database we will use is the ‘Adolescent Mental Health Data Platform’ (ADP), which is part of the ‘Secure Anonymised Information Linkage Databank’ (SAIL), describing the whole population of Wales, UK [13]. We have chosen this database as it most closely resembles the Cam-CHILD database; this was based on a review of the administrative datasets available in the UK which identified that SAIL included the most similar range and type of data in terms of contributing organisations, participants and character.

For the purposes of this study, we will link the datasets listed below within SAIL and restrict them to the last timepoint at which data was available from all datasets (i.e. last refresh date). Data will be in a structured format. In terms of recency, data are updated at varying points (often at least annually) but will not be refreshed in the lifetime of this project. Participants are assigned project-specific ‘ALFs’ by SAIL (Anonymous Linking Fields i.e. random identifiers) which are then used to link their records between datasets; participants may have more than one record within a dataset so records need to be compared to see whether multiple ALFs relate to multiple people or if it is a false duplication and in fact relates to the same person.

Baseline population data sources:-

- WDSD – Welsh Demographic Service:
  Provides records of births in Wales; will be used as a baseline of the whole Welsh population of young people.

- ADBE – Annual District Birth Extract:
  As a complementary dataset to the WDSD, the ADBE will be used to estimate the baseline population of Wales.

- ADDE – Annual District Death Extract:
  Records of all deaths of young people in the Welsh population and cause(s) of death if known.

- CENW – Census 2011 – Welsh Records:
  Complementary source of information on the baseline Welsh population. Also provides demographic information to be used for risk prediction models and to aid equity of algorithms e.g. ethnicity and nationality.

Social care sources:-

- LACW – Looked After Children:
  Provides records of looked-after children; will use to understand reason for looked-after status, type of support given, duration of looked-after status, safeguarding alerts, family make-up, onward referrals, progress updates, contacts with support team, diagnoses, outcome of looked-after care, reason for closure from looked-after children service etc. Sample size of dataset not publicly available; data covers 21/6/1999 - 31/3/2018.

- CRCS – Children Receiving Care and Support:
  Provides records of children receiving care and support; will use to understand reason for care and support status, type of support offered and provided, duration of care and support package, safeguarding alerts, family make-up, onward referrals, progress updates, contacts with support team, diagnoses, outcome of care and support contact, reason for closure from care and support service etc. Sample size of dataset = 16,000 young people; data covers 31/5/2009 - 31/5/2019.

Education source:-

- EDUW – Pre16 Education Attainment:
  Will be used to understand young people's attendance, attainment, next steps (e.g. college, university, employment, other), eligibility for free-school meals, exclusions (temporary or long-term), safeguarding alerts, onward referrals etc.

Healthcare sources:-

- WLGP – GP Primary Care- Audit:
  Provides records of GP contacts; will be used to understand health service contact, presenting problem(s), prescribed medication, diagnoses (mental and physical health), referrals to other services, safeguarding alerts etc.

- NCCH – National Community Child Health:
  Provides details of baby health checks and immunisations.

- MIDS – Maternity Indicators Dataset:

	As a complementary dataset to the 'National Community Child Health' dataset, providing details of baby health checks and immunisations. ● PEDW – Patient Episode Database for Wales: Provides details of hospital admissions; will be used to understand reason for admission/presenting problem(s), duration of visits, outcome, onward referrals, diagnoses, prescribed medication, safeguarding alerts etc. ● OPRD – Outpatient referrals from primary care: Provides records of outpatient referrals; will be used to understand the reason for referral/presenting problem(s), duration of problem, outcome, diagnoses given, medication prescribed, safeguarding alerts etc. ● OPDW – NHS Hospital Outpatients: Provides records of outpatient visits; will be used to understand presenting problem(s), duration, outcome, onward referrals, any diagnoses given, medication prescribed, safeguarding alerts etc. ● NHSO – NHS 111 Call data: Provides details of 111 calls; will be used to understand the reason for call/presenting problem(s), outcome of call etc. ● EDDS – Emergency Department Dataset: Provides details of emergency department visits; will be used to understand including reason for visit/presenting problem(s), duration, role of treating clinicians, treatment type, outcome, onward referrals, any diagnoses given, medication prescribed, safeguarding alerts etc. ● CCDS – Critical Care Dataset: Provides records of critical care contacts; will be used to understand reason for visit/presenting problem(s), duration, role of treating clinicians, treatment type, outcome, onward referrals, any diagnoses given, medication prescribed, safeguarding alerts etc. ● WRRS – Wales Results Reporting Service: Provides records of medical investigations. ● SMDS – Substance Misuse Dataset: Provides records of substance misuse contacts with young people.
Outcome	1. Period prevalence and distribution of mental and psychological health outcomes (including depression; anxiety; schizophrenia; bipolar; self-harm; substance misuse) among young people aged 0-17 years in Wales, UK; 2. Period prevalence and distribution of RFs associated with MH problems (including adverse childhood experiences; looked-after child status; safeguarding alerts; patterns of service use; educational attainment; environmental and community factors) among young people aged 0-17 years in Wales, UK; 3. Amount of unrecognised MH burden being managed by social care.

4. For machine learning algorithms, our outcome of interest will be the ability to predict the occurrence of a MH diagnosis occurring within a given time period.

Analytical strategy

Cohort construction and data linkage will be carried out in Structured Query Language (SQL). Due to staff expertise, data cleaning, exploration and analysis will be conducted in R, and machine learning methods and the prototype early identification tools will be developed in Python. Upon completion of the project, code will be made available through GitHub.

1. Cohort creation: We will create our cohort of interest [described in ‘Population’ section above]. Initially, anyone over the age of 20 years will be excluded to create a simpler database for linkage.

2. Linkage: ALFs (i.e. random unique IDs) will be used to link young people’s records between datasets.

3. Exploring Biases and Missingness: There are widely recognised limitations of using administrative datasets for the purpose of developing predictive models [17]. Our longer-term aim is to develop a tool that can identify young people at risk of MH problems prior to contact with MH or social care services, including as an ‘enriched’ population suitable for randomized interventional trials. As such, an important early step for our pilot will be to explore and characterise any systematic biases introduced into the database (a) as a result of failed linkage or inappropriate linkage [18-22], (b) as a result of exclusion in the dataset due to lack of contact with services or ethnicity [23-25], and (c) as a result of informed presence bias [26, 27]. Failed linkage will be estimated by examining the characteristics of matched versus unmatched records, looking for differences in factors such as age, gender, ethnicity, geographical location, site/clinical team, socio-economic status and health status/diagnoses. Once linked, we will assess data quality, for example by exploring the extent of missing data within records (e.g. proportion of health cases without diagnoses). We will also explore whether any variables (RFs) are particularly biased. The outputs of these studies will help inform which other datasets within the Cambridgeshire and Peterborough system could be included in the future build of the near-real-time database.

4. Database Characterisation: We will then move on to characterisation of the database starting with a description of the cohort at their index event (i.e. their first contact with either health or social care services). We will assess individual demographics (e.g. age, sex, socioeconomic status), the service to which individuals present (e.g. MH Trust, acute hospital Trust, social care team), and the outcome of any initial assessments (e.g. diagnoses or problem lists, measures of severity, outcomes of assessments). We will explore the prevalence and incidence of diagnoses and other identified needs (e.g. behavioural support) longitudinally within cases, and how these differ between the different health and social care settings, exploring how prevalent unidentified MH problems are in acute, community and social care settings and the time between the identification of initial symptoms and diagnosis of a MH problem.

5. Mapping RFs to Data Dictionaries: We have identified >190 RFs for young people’s MH problems through a rapid review of systematic reviews and reports, and discussions with clinicians

and professionals. These RFs are being prioritised through a Delphi study, whereby experts in young people's MH identify missing RFs and rank the importance of RFs in terms of predicting young people's MH problems. The prioritised list of RFs will be mapped to the linked ADP/SAIL database, identifying how they are operationalised across datasets (i.e. degree of harmonisation), which RFs can be measured in which services, and which can be measured in structured data only (vs which will be contained in unstructured data only and be a target for Natural Language Processing extraction in future projects). For RFs which are not directly measurable, we will identify proxy measures through literature review and discussion with professionals/clinicians (e.g. prescription of antidepressant in lieu of a recorded diagnosis of depression). Where possible, proxy measures will be sought from structured data.

6. Measuring MH Problems: Upon approval of data access, the ADP team have agreed to share a list of codes for 13 MH problems in people aged 10-24 (e.g. SNOMED and ICD codes). This code has been identified in the SAIL database so will be translatable for our purposes. We will use these identification codes to write programming code to extract information on MH problems within our cohort. With regards to analysis, the prevalence and distribution of MH problems will be measured by lower super output area (LSOA) code, analysing variation across geographical regions (NB: higher super output area will be used if LSOA would risk re-identification of individuals).

7. Measuring RFs for MH Problems: Using the RF to data dictionary mapping exercise in step 5, we will write code to extract information on RFs within our cohort. Next, the prevalence and distribution of RFs will be measured by lower super output area (LSOA) code and variation across geographical regions will be analysed (NB: higher super output area will be used if LSOA would risk re-identification of individuals).

8. Data Visualisation: Create heat maps demonstrating the prevalence and distribution of MH problems and RFs across Wales, UK.

9. Measuring Unrecognised MH Need in Social Care: Estimate how much unrecognised MH need is being met by social care services by calculating the proportion of young people with MH records who do not have MH problems documented in the structured data of their social care records.

10. Relationships Between RFs and MH Problems: We will describe the relationships between RFs and MH problems (e.g. causal relationships, moderating and mediating effects, bidirectionality).

11. Developing Predictive Algorithm Methods: There remains no consensus on the best approach to building risk prediction models in MH [28-30]. Preliminary characterisation of the variables in ADP/SAIL will inform our approach to data splitting, feature engineering and selection, model building, performance estimation and final model selection. Our aim is to understand which methods of risk prediction offer the best option in terms of accuracy, acceptability, explainability and ease of interpretation at the user interface (i.e. use by clinicians, professionals and the

public). As such, our methodology is purposely open here to allow testing and selection of appropriate models. We will compare risk prediction models which use traditional statistical methods (e.g. regression models) to those based on machine learning approaches (e.g. Random Forest and extreme gradient boosting/xgboost, with an explainable deep learning approach to derive rule-based findings), and a combination of the two approaches. This analysis will act as a pilot phase to establish the best methods to carry forward into analysis in the Cam-CHILD database. We will be assessing a range of different approaches to risk prediction, establishing the relative measures of success for each, to see which are most likely to be useful in predicting the risk of childhood MH problems. If shown to be sufficiently accurate, the risk prediction algorithms are intended to be an adjunct to clinical judgement, not a replacement for it, and will be used as a screening tool not a diagnostic one (e.g. a tool for use by social workers alongside their clinical expertise). As such, when measuring success of the pilot algorithms, we will favour having more false positives than false negatives (i.e. we prefer to reduce the likelihood of missing a genuine case of MH problems, even if it means incorrectly identifying some potential cases which turn out not to be true cases upon further investigation). We will use recall and F1 scores (i.e. a measure of precision and recall) as metrics of success, because the cost of false negatives outweighs the costs of false positives in our case. Where appropriate, we will measure the area under the receiver operating characteristic curve (AUC). Although we will test a variety of outcomes, our starting point will be to test a binary outcome – whether the risk prediction models can detect whether a young person will develop a MH problem or not (i.e. yes to any MH problems vs no to any MH problems). Subsequent evaluations will involve more complex outcomes (e.g. continuous outcomes such as prediction of the severity of MH problems).

12. Replication in Cam-CHILD, testing for generalisability: Analysis code above will be used to rapidly replicate our analyses in Cam-CHILD from April 2022, providing indication of generalisability of methods, findings and predictive models.

Ethics

Cam-CHILD database:-

- Ethics to access and link data across Cambridgeshire and Peterborough secured (REC ID: 20/EM/0299).
- Local information governance also agreed by all four data providers.

ADP/SAIL database:-

- After its initial construction, SAIL has not been required to seek additional consent to incorporate datasets arising from routine public service delivery because it is not a research activity per se and data is accessed by researchers in an anonymised form [31]. Formal permission for data usage from each participating organisation (i.e. data providers) is granted via their Caldicott Guardians and Information Governance structures [32].

- Project-specific approval was sought and reviewed by an Information Governance Review Panel (IGRP) process. Access to most datasets was approved on 11/10/21. Approval for three restricted datasets (LACW, EDUW, CRCS) is pending, and will be reviewed by the Welsh Government. Approval for use of the Census data is being sought. For approved datasets, we will be able to commence data exploration once data has been provisioned by SAIL. Project number: 1336; project name: 'Measuring mental health problems across an integrated child health system, together with associated risk factors: towards developing early identification models for childhood mental health problems'.

Risks

Risk	Mitigation
Delays in receiving data from ADP/SAIL.	We have given conservative estimates of time to complete subsequent tasks and allowed contingency time at the end of the project. In addition, we can begin attempting to map RFs whilst awaiting access by utilising metadata on the SAIL website and exploring variables in publications arising from SAIL.
Data cleaning takes longer than expected. From discussion with professionals who have used this database and similar datasets, we anticipate that data cleaning is likely to be a big task (more so even than data analysis).	To mitigate this, we have been conservative in our estimates of time required for cleaning and ensured we have appropriate support. More time has been allocated for data cleaning and upfront work in the project. We will also draw on the expertise of people who are familiar with the database and similar ones. Where available, we will use open-source code to expedite data cleaning. Where possible, we will also fund additional support from the ADP team to aid with data linkage as they have expertise in this database.
Poor data quality leads to difficulties with analysis and build of risk prediction tools.	At the initial point of data inclusion in the ADP, this risk is managed by the below (from: https://www.adolescentmentalhealth.uk/Platform#SecureResearchEnvironment) [3]: "Any new data will pass through our Data Quality Analysis tool first to ensure that you produce precision data to reduce errors that would otherwise affect your research." At the point of our linkage and analysis, this risk will be mitigated by: drawing on expertise in data management, statistics and ADP data; reading articles on how other researchers have approached this issue in SAIL; exploring biases which may be introduced through failed linkages; making appropriately cautious claims and inferences from

	data based on recognition of any data quality issues.
Analysis takes longer to complete than expected.	We have included access to data for three researchers, meaning there is more flexibility in our ability to deliver on time. Moreover, data access is based on 2-5 researchers so we have two unfilled access places should we need additional support. Also analyses can be carried out concurrently if required.
Delays due to impact of COVID-19.	Due to the nature of the project (desk based with little requirement for face-to-face contact), this project has low probability of negative impact due to COVID-19. We have established good remote working practices with social care experts, meaning we have good access to domain expertise.
Cam-CHILD database is not ready for replication of analysis in April 2022.	We will receive this data in Sept/Oct 2021, enabling work to clean the data to start. We have existing funding to support this work, providing 6/7 months to prepare the database, which is a conservative estimate of the time required. If needed we will draw on resources within the university to increase the capacity of the Cam-CHILD team.

Registration

This project was registered with the Open Science Framework (OSF) on 19/11/2021. Registration link: https://osf.io/ukn6b

Data Protection

The data used for the research is administrative data collected in the course of social care, healthcare, local authority services, and government carrying out their day-to-day duties. No further data will be collected for the purpose of this research. Our team will not store any raw data; all data will be accessed via a Virtual Desktop Infrastructure (VDI).

Before data is incorporated into SAIL Databank, data providers (e.g. GPs, hospitals, government departments etc) separate their datasets into two parts: a demographic component and a content component. Content information is sent directly to SAIL, whilst demographic information is processed by a National Health Service based Trusted Third Party in the NHS Wales Informatics Service (NWIS). Only the ALFs with some minimal demographic data (including gender, week of birth and area of residence to 1500 head of population) are sent to SAIL for recombination with the content data. Because SAIL does not have access to or control over patient identifiable data, they do not become a data controller. [31]

The Health Information Research Unit (HIRU) is the data custodian, but there is shared control over access and use of the data through an IGRP (Information Governance Review

Panel). Applications to use SAIL data are reviewed by an IGRP, which is made up to include representatives from Informing Healthcare, the National Research Ethics Service, the National Public Health Service for Wales, the British Medical Association and Involving People. [32]

According to Jones et al. (2019) “SAIL is not required to seek additional consent to incorporate datasets arising from routine public service delivery. This is because it is not a research activity per se and data accessed by researchers are in anonymised form. In accordance with the GDPR, [they] provide privacy notices on behalf of data providers (in places such as in General Practice surgeries). These inform members of the public of data use, and individuals are able to opt-out of their data being provided to SAIL by informing their GP. The opt-out is enacted between the data provider and NWIS: in practice [SAIL] have had less than 0.025% of the population make this request to date”. [31]

What Works for Children’s Social Care (WWCSC) is not a data controller or processor for any data in relation to this project.

Principles of GDPR [34]:

Principle (a): Lawfulness, fairness and transparency

1. Lawfulness: The lawful basis for processing this data is to perform a task carried out in the public interest. [35]
2. Fairness: According to ICO’s guidance, fairness means “you should only handle personal data in ways that people would reasonably expect and not use it in ways that have unjustified adverse effects on them” [36]. We consider that young people and families reasonably expect research to be carried out to aid service improvement and innovation. Furthermore, consultation with our expert-by-experience reference group (aka EbERG; group includes those with lived experience of social care) has indicated that using routinely-collected social care, healthcare and education data to identify MH problems early, explore MH burden and identify service provision gaps is a valuable and important use. Moreover, some members noted that such tools could have improved their patient journey by facilitating smoother pathways between services.
3. Transparency: According to ICO’s guidance [36], “Transparent processing is about being clear, open and honest with people from the start about who you are, and how and why you use their personal data…You must ensure that you tell individuals about your processing in a way that is easily accessible and easy to understand. You must use clear and plain language.” This includes situations of ‘invisible processing’ such as ours where there is no direct relationship with the individual and their personal data is collected from another source. To promote transparency and understanding, we will be publicising the Cam-CHILD project widely, in particular through networks of people with lived experience of MH problems. For this, we are supported by the Anna Freud National Centre for Children and Families, who will assist with disseminating the research and its outputs in an understandable format. The EbERG will also be consulted on an ongoing basis with regards to how to ensure early identification tools are acceptable to those who will interact with them.

Principle (b): Purpose Limitation
Data will be used to answer the questions outlined above and contribute to answering the overall research questions for the Cam-CHILD project. They will not be used for any other purpose.

Principle (c): Data Minimisation

Due to the nature of machine learning approaches, significant amounts of data are required to build accurate models with reduced bias. We have requested data that is adequate, relevant and limited to what is necessary for developing predictive models, and to fulfil the purpose of this project overall.

In terms of special categories of personal data, we have requested data from SAIL on ethnicity, nationality, and language of young people (via linked ONS data). This is because we want to ensure we have considered demographic information in our MH risk prediction models and also to ensure the algorithms we develop are not introducing or exacerbating health or social inequalities or bias. We also want to know how demographics relate to MH risk and how predictive our models are for different groups, especially those who may be considered minority groups, to ensure we are not building biases into the algorithms.

Principle (d): Accuracy
The data quality is checked when it enters SAIL and will undergo checks by our team.

Principle (e): Storage limitation
Our team will not store any raw data; all data will be accessed via a Virtual Desktop Infrastructure (VDI). Furthermore, we will only access the data for the period of time necessary to do the research and appropriate quality assurance.

Principle f): Integrity and confidentiality (Security)
Identifiable data is never released from ADP servers, and is instead accessed using a Virtual Desktop Infrastructure (VDI). When researchers would like data to generate figures, images or tables, the data is reviewed for anything potentially identifiable before being released. Single rows of data cannot be released and instead aggregate data will be reported to maintain anonymity (for example, when mapping MH problems by lower super output area). When anonymity may be compromised by data granularity, data will be aggregated further and LSOA may not be used.

Principle (g): Accountability principle
According to ICO's guidance, this refers to "Taking responsibility for what you do with personal data, and demonstrating the steps you have taken to protect people's rights". Both ADP/SAIL and our team will ensure the protection of people's personal data, for example, by only releasing aggregated results and actively engaging with people with lived experience of MH problems to ensure we are meeting public expectations. Researchers accessing SAIL databank are also required to complete appropriate GDPR training and comply with this. Furthermore, by publishing this protocol openly, this adds another channel of accountability.

Personnel

Team Member	Role(s) in Project	Job Title	Organisational Affiliation
*Dr. Anna Moore	PI – overall lead and study design. Supervision of Katherine Parkin. Statistical analysis. Report write up. Application in Cam-CHILD.	NIHR Clinical Lecturer Child Psychiatry	University of Cambridge
Prof. Tamsin Ford	Supervision for Dr. Anna Moore.	Professor of Child Psychiatry	University of Cambridge
*Katherine Parkin	Draft code for data extraction from ADP/SAIL, data and code management, data cleaning, statistical analysis with supervision, liaison with social workers and clinicians as required.	PhD student	University of Cambridge
*Dr. Efthalia Massou	Statistician, will provide supervision, support statistical plan, help with analysis.	Research Associate in Primary Care Methodology Unit	University of Cambridge
Prof. Pietro Liò	Supervision for AI/ML aspects.	Professor in Computer Science and Technology, AI methodologist.	University of Cambridge

*These team members will access the ADP/SAIL data.

Timeline

Completion Dates	Activity	Staff Responsible/Leading
Oct 2021	Receive data from ADP/SAIL	ADP/SAIL
31st Oct 2021	Map RFs to ADP/SAIL data dictionaries	AM, KP & EM
15th Nov 2021	Develop code to extract data on RFs and MH problems	AM, KP & EM
31st Dec 2021	Link and clean data	AM, KP & EM (supported by ADP team)
15th Mar 2022	Complete measurement of prevalence and distribution of MH problems and associated RFs	AM, KP & EM
15th Mar 2022	Analyse relationships between MH problems and RFs	AM, KP & EM

15th Mar 2022	Measure unidentified MH needs in social care	AM, KP & EM
15th Jun 2022	Replicate analyses in Cam-CHILD database	All (led by AM)
31st Aug 2022	Preliminary methods for risk prediction tools developed	AM, KP, EM & PL
Autumn 2022	Dissemination complete: submission of paper, conference presentation, blogs and dissemination of findings	AM & KP

References

1 Ford T, Hamilton H, Goodman R, Meltzer H. Service contacts among the children participating in the British child and adolescent mental health surveys. *Child and Adolescent Mental Health*. 2005 Feb;10(1):2-9.

2 Department for Education. *Characteristics of children in need 2020*. 2020 [cited 2021 July 7]. Available from: https://explore-education-statistics.service.gov.uk/find-statistics/characteristics-of-children-in-need.

3 Department for Education. *Children looked after in England including adoptions.* [Internet]. 2021 [cited 2021 April 7]. Available from: https://explore-education-statistics.service.gov.uk/find-statistics/children-looked-after-in-england-including-adoptions/2020.

4 Allen G. Early intervention: the next steps, an independent report to Her Majesty's government by Graham Allen MP. The Stationery Office. 2011.

5 Maguire A, Ross E, O'Hagan D, O'Reilly D. RF12 Suicide ideation and mortality risk: population wide data linkage study. *Journal of Epidemiology and Community Health*. 2019;73(Suppl 1):A60-A: 10.1136/jech-2019-SSMabstracts.127

6 Berridge D, Luke N, Sebba J, Strand S, Cartwright M, Staples E, McGrath-Lone L, Ward J, O'Higgins A. Children in need and children in care: Educational attainment and progress. University of Bristol, University of Oxford. 2020 Apr 27:2020-05.

7 What Works for Children's Social Care. *Study Review: Mental Health Care Interventions for Children Looked After*. 2016 [cited 2021 October 15]. Available from: https://whatworks-csc.org.uk/evidence/evidence-store/intervention/mental-health-care-interventions-for-children-looked-after/

8 The Child Safeguarding Practice Review Panel. Annual Report 2020: Patterns in practice, key messages and 2021 work programme. 2021. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/984767/The_Child_Safeguarding_Annual_Report_2020.pdf

9 Care Leavers' Association. Caring for Better Health: An investigation into the health needs of care leavers. 2017. London: CLA. Available from: https://www.careleavers.com/wp-content/uploads/2017/12/Caring-for-Better-Health-Final-Report.pdf

10 Fong H-f, French B, Rubin D, Wood JN. Mental health services for children and caregivers remaining at home after suspected maltreatment. *Children and Youth Services Review.* 2015;58:50-9.

11 Sanders R. Care experienced children and young people's mental health. [Internet]. 2020 [cited 2021 October 17]. Available from: https://www.iriss.org.uk/resources/esss-outlines/care-experienced-children-and-young-peoples-mental-health

12 Department of Health and Social Care and Department for Education. *Transforming children and young people's mental health provision: A green paper.* Assets Publishing website. 2017.

13 Lyons RA, Jones KH, John G, Brooks CJ, Verplancke JP, Ford DV, Brown G, Leake K. The SAIL databank: linking multiple health and social care datasets. *BMC medical informatics and decision making*. 2009 Dec;9(1):1-8.

14 Kreps GL. Stigma and the reluctance to address mental health issues in minority communities. *Journal of Family Strengths.* 2017;17(1):3-13.

15 McFadden A, Siebelt L, Gavine A, Atkin K, Bell K, Innes N, Jones H, Jackson C, Haggi H, MacGillivray S. Gypsy, Roma and Traveller access to and engagement with health services: a systematic review. *The European Journal of Public Health*. 2018 Feb 1;28(1):74-7.

16 Whitley R, Kirmayer LJ, Groleau D. Understanding immigrants' reluctance to use mental health services: a qualitative study from Montreal. *The Canadian Journal of Psychiatry.* 2006 Mar;51(4):205-9.

17 Lusa L. Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models. *BMC bioinformatics*. 2015 Dec;16(1):1-0.

18 Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. *Annals of Human Biology*. 2020 Feb 17;47(2):218-26.

19 Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, Goldstein H. Challenges in administrative data linkage for research. *Big data & society*. 2017 Dec;4(2):2053951717745678.

20 Kharrazi H, Wang C, Scharfstein D. Prospective EHR-based clinical trials: the challenge of missing data. *J Gen Intern Med*. 2014 May;29(7):976–8.

21 Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM, RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS medicine.* 2015 Oct 6;12(10):e1001885.

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

23 Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *The Lancet Digital Health.* 2021 Mar 4;3(4):e260-5.

24 Mc Grath-Lone L, Libuy N, Etoori D, Blackburn R, Gilbert R, Harron K. Ethnic bias in data linkage. *The Lancet Digital Health.* 2021 Jun 1;3(6):e339.

25 Knight HE, Deeny SR, Dreyer K, Engmann J, Mackintosh M, Raza S, Stafford M, Tesfaye R, Steventon A. Challenging racism in the use of health data. *The Lancet Digital Health*. 2021 Mar 1;3(3):e144-6.

26 Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American journal of epidemiology.* 2016 Dec 1;184(11):847-55.

27 Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC medical informatics and decision making.* 2014 Dec;14(1):1-9.

28 Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, Liu X, He Z. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*. 2020 Jul;27(7):1173-85.

29 Bernert RA, Hilberg AM, Melia R, Kim JP, Shah NH, Abnousi F. Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. *International journal of environmental research and public health*. 2020 Jan;17(16):5929.

30 Liu GD, Li YC, Zhang W, Zhang L. A brief review of artificial intelligence applications and algorithms for psychiatric disorders. *Engineering*. 2020 Apr 1;6(4):462-7.

31 Jones KH, Ford DV, Thompson S, Lyons RA. A profile of the SAIL databank on the UK secure research platform. *International Journal of Population Data Science*. 2019;4(2).

32 Ford DV, Jones KH, Verplancke JP, Lyons RA, John G, Brown G, Brooks CJ, Thompson S, Bodger O, Couch T, Leake K. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC health services research*. 2009 Dec;9(1):1-2.

33 Adolescent Mental Health Data Platform. Secure Research Environment. [Internet]. 2021 [cited 2021 September 12]. Available from: https://www.adolescentmentalhealth.uk/Platform#SecureResearchEnvironment

34 Information Commissioner's Office. The Principles. [Internet]. 2021 [cited 2021 September 12]. Available from: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/

35 European Parliament. Regulation (EU) 2016/679 of the European Parliament and of the Council. Chapter 2, Article 6. [Internet]. 2016 [cited 2021 September 12]. Available from: https://www.legislation.gov.uk/eur/2016/679/article/6

36 Information Commissioner’s Office. Principle (a): Lawfulness, fairness and transparency. [Internet]. 2021 [cited 2021 September 12]. Available from: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/lawfulness-fairness-and-transparency/

Bibliography

This protocol was also informed by the following source:

Clayton V, Bogiatzis Gibbons D, Sanders M. *Pilots of predictive analytics in children's social care: Protocol*. 2019 [cited 2021 September 12]. Available from: https://osf.io/jwtf4/