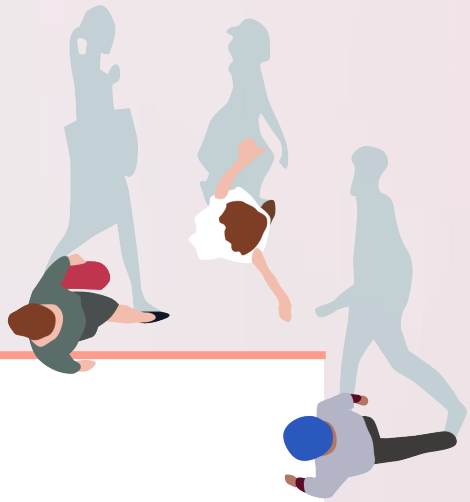
What Works *for* Children's Social Care

EARLY INTERVENTION FOUNDATION

Coming together as What Works for Early Intervention & Children's Social Care



TOWARDS EARLY IDENTIFICATION OF MENTAL HEALTH PROBLEMS IN CHILDREN'S SOCIAL CARE

February 2023

What Works for Children's Social Care

EARLY INTERVENTION FOUNDATION

Coming together as What Works for Early Intervention & Children's Social Care

Acknowledgements

This study makes use of anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. We would like to acknowledge all the data providers who make anonymised data available for research. This work was supported by the Adolescent Mental Health Data Platform (ADP). The ADP is funded by MQ Mental Health Research Charity (Grant Reference MQBF/3 ADP). ADP and the author(s) would like to acknowledge the data providers who supplied the datasets enabling this research study. The views expressed are entirely those of the authors and should not be assumed to be the same as those of ADP or MQ Mental Health Research Charity. We would also like to thank Professor Pietro Liò, Emma Rocheteau, Dr Angela Wood, and Professor Zoe Kourtzi for providing supervision and guidance for the machine learning aspects of this project. Finally, we would like to thank our funders, without whom this work would not have been possible.

Authors

Katherine Parkin, Ryan Crowley, Efthalia Massou, Marcos Del Pozo Baños, Yasmin Friedmann, Ann John, Anna Moore

Funding

This research was funded as part of the WWCSC Spark Grant Scheme.

All research at the Department of Psychiatry in the University of Cambridge is supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) and NIHR Applied Research Collaboration East of England. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Katherine Parkin is funded by the National Institute for Health and Care Research (NIHR) School for Public Health Research (SPHR) (Grant Reference Number PD-SPH-2015) and the NIHR Applied Research Collaboration (ARC) East of England.

Dr Anna Moore is funded through an NIHR Clinical Lectureship funded by Anna Freud National Centre for Children and Families (AFC). The Delphi Study, which provided foundational work for this project, was funded by MRC Adolescent Engagement Awards MR/T046430/1. Dr Moore also holds grants from the Alan Turing Institute and UKRI/DARE UK. Data linkage within SAIL was carried out by Dr Yasmin Friedmann at the Adolescent Mental Health Data Platform (ADP), and was funded by Cambridgeshire and Peterborough NHS Foundation Trust (CPFT).

Dr Marcos Del Pozo Baños, Dr Yasmin Friedmann and Professor Ann John are funded through the ADP, which is funded by MQ Mental Health Research Charity (Grant Reference MQBF/3 ADP). The views expressed are entirely those of the authors and should not be assumed to be the same as those of ADP or MQ Mental Health Research Charity.

The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care, the AFC, the MRC, the Alan Turing Institute, the UKRI, the ADP or MQ Mental Health Research Charity.

No competing interests declared.

About What Works for Early Intervention and Children's Social Care

What Works for Children's Social Care (WWCSC) and the Early Intervention Foundation (EIF) are merging. The new organisation is operating initially under the working name of What Works for Early Intervention and Children's Social Care.

Our new single What Works centre will cover the full range of support for children and families from preventative approaches, early intervention and targeted support for those at risk of poor outcomes, through to support for children with a social worker, children in care and care leavers.

To find out more visit our website at: **www.whatworks-csc.org.uk**

About the Timely Project (Department of Psychiatry, University of Cambridge)

The Timely project, led by Dr Anna Moore, is seeking to develop digital tools to identify young people's mental health problems early. The project is using linked health, social care, education and genetic data, and it explores the value of machine learning and artificial intelligence approaches.

Spark Grant Scheme

This research was funded as part of the WWCSC Spark Grant Scheme. The purpose of the scheme is to fund new and innovative research in children's social care, conducted by researchers who may struggle to secure funding through other routes, particularly Early Career Researchers (ECRs) and/or researchers from underrepresented, minoritised groups. This work is an important part of our mission to develop capacity within the research community and generate high-quality evidence in children's social care.

CONTENTS

Executive summary

Background

Due to poor integration of data held about young people, there remain significant problems in identifying young people with mental health problems in social care settings. Young people and families can suffer for prolonged periods without suitable mental health support (DfE, 2020, 2021). Almost all young people with social care contact are likely to experience some kind of mental health problem, yet only a small proportion of them are thought to have a formal diagnosis and even less receive treatment (Berridge et al., 2020; Care Leavers' Association, 2017; The Child Safeguarding Practice Review Panel, 2021). Failure to identify risk factors and mental health-associated problems early can delay treatment and lead to limited interventions failing to address significant causes of a young person's difficulties (Allen, 2011; DHSC & DfE, 2018). If accurate early identification tools could be developed, this would help young people in children's social care receive more timely support.

Routinely collected data from health, social care and education settings contain broad-ranging information which can be used to understand an individual's exposure to risk factors for mental health problems. Thus, using these data may facilitate the prediction of mental health outcomes. Machine learning methods could benefit from the large amount of data available in routinely collected datasets; these algorithms can use this information to learn from existing data and discover patterns which are then used to predict the outcome of future observations. These machine learning approaches could supplement the standard statistical approaches which are often used. We can then compare their performance and investigate the relative benefits of both methods (traditional statistical approaches vs machine learning approaches), in terms of the amount of predictive utility they offer, their interpretability, and the relative challenges associated with their implementation.

Previous machine learning models have not been able to reach the performance needed for clinical use in children's social care settings (Clayton et al., 2020). In part, this may be explained by the datasets used to build these models to date. The evidence describing the mechanisms underlying mental ill health is rapidly evolving to reveal the role of biological factors (e.g. physical health, immunology, inflammation and genetics). These interact with early life experiences and the environment to confer resilience and susceptibility to mental health problems. Therefore, given this multi-factorial nature of mental ill health, we hypothesise that building accurate models to identify mental health problems will require access to large, representative datasets of multi-domain data that reflect this broad range of bio–psycho–social factors. In previous work (not yet published), we developed a framework of risk factors for mental health problems based on Bronfenbrenner and Morris's 2006 bioecological model (Bronfenbrenner & Morris, 2006), with emphasis on identifying risk factors relevant to underserved populations. This work resulted in the identification of 287 risk factors which we grouped into a framework of eight domains. In this report, we present work from the next stage of the study, in which we explore the utility of routinely collected datasets for predicting young people's mental health problems. To do this, we created and characterised a linked multi-agency database (containing datasets from health, social care and education) relating to most children in Wales and including the broad range of variables

identified in our earlier work. We used this linked database to measure the prevalence of mental health problems within the cohort. This database was then used to explore various machine learning methods to identify mental health problems in children in social care settings.

Objectives and research questions

In this study, we aimed to create a linked database of health, social care and education data in order to measure childhood mental health problems and their associated risk factors. A linked database is useful for this purpose as it allows mental health problems and associated risk factors to be measured from different agencies with which young people come into contact, as opposed to limiting measurement to a single agency (such as GP or A&E); this should allow for more comprehensive and holistic measurement of mental health problems and their associated risk factors.

In addition to this measurement, we aimed to develop prototype models for early identification of mental health problems of young people in social care settings. This work was carried out within the Secure Anonymised Information Linkage (SAIL) Databank, but will also inform the development of a similar linked database in Cambridgeshire and Peterborough, known as CADRE (Child and Adolescent Data REsource; [formerly known as Cam-CHILD]).

The study research questions were:

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

Design

For the measurement of mental health problems and associated risk factors, a retrospective cohort study design was used, with cross-sectional analysis. For machine learning approaches, the same cohort study and particular elements of this (i.e. site-level data) were used, with the data being split into training, test and validation sets.

Our cohort of young people was defined as anyone who was aged 0–17 years in the period between 1 January 2013 and 31 March 2020; all retrospective and subsequent data for these individuals were included even if it fell outside this time period. The final cohort consisted of 1.1 million young people in Wales, of which 46,704 had social care data and were thus used in sub-sample analysis for early identification model prototyping. Though the

overall cohort comprises 1.1 million young people, sample sizes differ quite substantially between datasets (as shown below in Table 3.2).

Findings

When measured in the Welsh GP dataset (WLGP), 14.85% of our cohort had at least one mental or psychological health condition of interest, with mood disorders being most common (12.96%) and severe mental illness (SMI) such as schizophrenia and bipolar being least common (0.11%). When measured in the Patient Episode Dataset for Wales (PEDW), the prevalence of any mental health condition of interest was 4.78%, with mood disorders still being most common (2.73%) and SMI being least common (0.13%). In the rest of the datasets we used, the prevalence of any mental health conditions of interest was between <0.00% and 1.33%.

With regards to the measurement of risk factors, we found risk factors fell on a spectrum of measurability, ranging from “directly measurable” to “derivable” to “measurable by proxy” (defined in Table 3.3). We focused our efforts on the former two, and found important factors associated with childhood mental health problems were spread across different data sources, rather than being confined to any one particular database. Of 287 risk factors of interest, 101 (35.19%) were measurable, of which 48 (16.72%) were directly measurable and 53 (18.47%) were derivable. Of 101 risk factors of interest relating specifically to underserved populations, 37 (36.63%) were measurable; this was broken down into 26 (25.74%) which were directly measurable and 11 (10.89%) which were derivable.

For the prototype early identification model in social care settings, we developed simple statistical models and both basic Neural Network models with Rectified Linear Unit (ReLU) activations and Graph Neural Networks (GNNs). The best-performing GNN model achieved an AUROC (Area Under the Receiver Operating Characteristics Curve) of .815 and the best-performing Neural Network achieved an AUROC of .800. These results indicate that the GNN approach may provide a promising method for identifying young people with a mental health diagnosis. However, greater accuracy and further validation is required prior to considering clinical implementation. In comparison, standard logistic regression models achieved an AUROC of .803.

Conclusions

This work in the SAIL Databank demonstrated that it was possible to link together multi-agency data from social care, health and education settings. With this linked data, we were then able to measure the prevalence of different mental health conditions and their associated risk factors. Due to creating a linked multi-agency database, we were able to measure different bio–psycho–social risk factors which would not have been measurable in single-agency data. We could then include these in prototype early identification models. Though these models’ performance was not sufficient for clinical use, they provide a solid foundation to improve on. Mental health problems have bio–psycho–social causes and correlates; thus, if we are to build accurate and implementable early identification tools, bio–psycho–social databases from routine sources are likely to be required. This project

highlights the worth of bringing together rich data from different organisations with which young people interact.

In summary, linkage of multi-agency data offers a promising way of developing early identification tools because early warning signs for mental health problems which may be missed in single-agency data can be combined, leading to a stronger signal for detecting developing problems. Early identification of potential problems means that young people and their families can be offered more timely and proportionate support, instead of waiting in distress for problems to worsen and meet service thresholds. Furthermore, as more robust early identification tools are developed, staff in contexts such as social care can use them as an adjunct for decision-making to help them identify young people who may have additional needs and to support smoother care pathways for young people and their families.

Ongoing work

This study is currently ongoing. As such, in this report, we present the findings to date. Thus far, we have: gained approval to access all 18 desired databases; linked 18 databases in SAIL; characterised our 18 databases of interest; characterised our cohort of interest; mapped 287 risk factors to SAIL metadata and explored their measurability (**Research Question 1**); measured mental health problems in health datasets (**Research Question 1**); explored relationships between risk factors and mental health outcomes (**Research Questions 4 and 5**); and developed an early prototype of a risk prediction tool for social care settings (**Research Question 5**).

We have successfully applied to extend our access to the SAIL Databank and will continue our analysis in the linked database we have created as part of this work. Ongoing work involves: measuring the prevalence and distribution of risk factors for mental health problems (**Research Question 2**); measuring the prevalence of mental health problems in the health datasets when linked together (**Research Question 2**); measuring the distribution of mental health problems by region (**Research Question 2**); measuring unrecognised mental health need in social care (**Research Question 3**); improving the accuracy of the protype risk prediction tool (before it could be considered for clinical use) (**Research Question 5**); and replicating this work in the CADRE database (**Research Question 6**).

1. Introduction

Background and problem statement

There are high levels of mental health need in children's social care settings (Berridge et al., 2020; DfE, 2020, 2021; Maguire et al., 2019). 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 mental health need in this population. Access to childhood mental health support can be challenging, and there are even more barriers to access for young people within children's social care settings (What Works for Children's Social Care, 2016). It is important to provide suitable mental health 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 mental health needs in social care settings. Moreover, young people in these settings have distinct mental health needs (Care Leavers' Association, 2017; The Child Safeguarding Practice Review Panel, 2021), and there is some evidence that standard mental health interventions may be harmful for some young people with a history of social care contact, for example, looked-after children (Fong et al., 2015). As such, it is critically important to provide risk factor-informed interventions to this population (for example, specific trauma-informed and non-stigmatising interventions). At present, without this approach, outcomes for young people with mental health 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 (i.e. Not in Education, Employment or Training), all of which can lead to poor long-term health and social outcomes (Sanders, 2020).

In summary, without an effective means of early identification, young people and their families can suffer for prolonged periods without suitable mental health support (DfE, 2020, 2021). Furthermore, a failure to identify risk factors and mental health-associated problems early can delay treatment and lead to limited interventions failing to address significant causes of a young person's difficulties (Allen, 2011; DHSC & DfE, 2018). If accurate early identification tools could be developed, this could help young people in children's social care receive more timely support. Machine learning methods offer one potential way to learn from existing data on risk factors for young people's mental health problems in order to build effective predictive models for early identification of mental health problems. Previous machine learning models have not been able to reach the performance needed for clinical use in children's social care settings (Clayton et al., 2020). In order to build accurate risk prediction models suitable for clinical implementation, we hypothesise that an approach using linked, multi-agency data with a large number of observations is required, effectively linking risk factor data from social care, education and healthcare datasets.

To address the aforementioned problems, we suggest that we need to:

- Accurately understand the prevalence and distribution of mental health problems and associated risk factors in social care settings across different geographical regions

- Understand the specific relationships between risk factors and mental health outcomes
- 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
- Develop reliable early identification tools, which do not rely on already overstretched Child and Adolescent Mental Health Services (CAMHS).

Our hope is that this will lead to:

- Quicker access to assessment and intervention
- Enablement of research into interventions for young people with mental health problems in social care
- Clarity for social workers about young people who are challenging to diagnose and signpost
- Facilitation of conversations about access to mental health services. In turn, this will improve outcomes and experiences for young people and their families through better integration and access to mental health services.

Study aims

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 mental health problems and risk factors within multi-agency data
    b. Develop methods to map the prevalence and distribution of mental health problems and associated risk factors in multi-agency data
    c. Estimate unidentified mental health need within social care.
2. Explore relationships between exposure to risk factors and mental health outcomes.
3. Explore the best methods for developing accurate and usable child and adolescent mental health risk prediction algorithms.
4. Begin applying these methods to the CADRE database, in order to test validity of the database and generalisability of the risk prediction algorithms.

Research questions

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

6. Can findings and methods be replicated across databases (i.e. translated to CADRE)?

2. Methods

Data sources

In an earlier database review exercise, we compared a number of administrative datasets available in the UK. Through this analysis, we identified the Secure Anonymised Information Linkage Databank (SAIL) as the most characteristically similar data source to CADRE (Child and Adolescent Data REsource), the database we are developing in Cambridgeshire and Peterborough which will contain administrative data for young people aged 0 to 17 years from health, social care and education, along with a research database of genetic data. Our aim is for CADRE to be usable by clinicians in a real-time fashion to aid clinical decision-making, and a de-identified version will be accessible for approved researchers and research projects. Similarities between SAIL and CADRE include the range and type of data in terms of contributing organisations, participants and character. SAIL Databank is a national data safe haven of de-identified datasets relating to the population of Wales. SAIL operates on the UK Secure Research Platform (UKSeRP). Researchers can apply to access SAIL data in an anonymised form via this secure research environment (Jones et al., 2019; Lyons et al., 2009).

We identified 18 SAIL databases which would be relevant to our project. These databases were administrative or routinely collected and broadly pertained to demographics (including births and deaths), social care, education and health. Specifically, the databases were:

- Welsh Demographic Service (WDSD)
- Annual District Birth Extract (ADBE)
- Annual District Death Extract (ADDE)
- Child in Need Dataset – Wales (CINW)
- Children Receiving Care and Support (CRCS)
- Looked After Children – Wales (LACW)
- Pre-16 Education Attainment (EDUW)
- GP Primary Care – Audit (WLGP)
- National Community Child Health (NCCH)
- Maternity Indicators Dataset (MIDS)
- Patient Episode Database for Wales (PEDW)
- Outpatient Referrals from Primary Care (OPRD)
- NHS Hospital Outpatients (OPDW)
- NHS 111 Call Data (NHSO)
- Emergency Department Dataset (EDDS)
- Critical Care Dataset (CCDS)
- Wales Results Reporting Service (WRRS)
- Substance Misuse Dataset (SMDS).

Data management and pre-processing

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 was collected for the purpose of this research. Our team did not store any raw data; all data was accessed via a Virtual Desktop Infrastructure (VDI).

Before data is incorporated into the 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, while demographic information is processed by a National Health Service-based Trusted Third Party in the NHS Wales Informatics Service (NWIS). Only Anonymised Linkage Fields (ALFs) with some minimal demographic data (including gender, week of birth and general area of residence) 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 (Jones et al., 2019). 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) (Ford et al., 2009).

All data are treated in accordance with the Data Protection Act 2018. 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."*

For the most part, the data provisioned to researchers by SAIL is very similar to the original unprocessed data available to clinicians and professionals in routine practice, but with unstructured (i.e. free-text notes) omitted. According to the SAIL support team, there is some minimal data processing that happens for the WDSD, PEDW and WLGP (for example, with the WLGP, the GP registration records for individuals are simplified and anomalies such as short gaps and overlaps are resolved by applying a set of rules; [see Thayer et al., 2020 for details]), but other than this, the SAIL team tends to provide data to researchers as it is received from data providers. The SAIL team can make recommendations to researchers about data cleaning and processing, but generally the process of data cleaning is left to researchers as each project will have unique requirements (personal communication, 13 December 2022). More detailed information about each dataset, including time lags for incorporation into SAIL, can be found on the HDRUK website by searching per dataset.[1]

[1] See https://www.healthdatagateway.org/.

For our project specifically, during data linkage, Dr Friedmann also included a generated variable, the Welsh Index of Multiple Deprivation 2011 (or WIMD), in the WDSD for us; this variable was created by the ADP team. The WIMD 2011 variable is an official measure of deprivation in small areas of Wales (defined as containing approximately 1,500 individuals), based on employment opportunities, income, education, health, community safety, geographical access to services, housing and the physical environment.

Data access and information governance

We completed the application process to request access to the 18 aforementioned databases. Our application was reviewed by the internal and external Information Governance Review Panel (IGRP). The IGRP includes representatives from Informing Healthcare, the National Research Ethics Service, the National Public Health Service for Wales, the British Medical Association and Involving People (Ford et al., 2009).

Following this process, we were granted approval to access all 18 databases within the SAIL Databank. The SAIL data analysts then set up our gateway access and provisioned our data, including setting up our project views in the SAIL gateway and creating a cohort based on our age and date specifications.

A note on reporting small-value results

To mitigate against re-identification of individuals, SAIL stipulates that any reported results must include a minimum of five individuals (and a minimum of ten individuals for data from the Office for National Statistics census, which was not included in the present project). Results smaller than this must be aggregated. Where this occurs, we have indicated it with the symbol ✧ and a note explaining this.

Data linkage

Data linkage was carried out by Dr Friedmann, a Senior Research Data Scientist from the Adolescent Mental Health Data Platform (ADP), a project focusing on children and adolescents within the SAIL Databank. Data was linked on an individual level via the ADP. The data linkage process matches individuals' records between datasets and facilitates de-duplication of individuals' records.

Individuals within the SAIL datasets are assigned a unique Anonymised Linkage Field (ALF) that replaces any identifiable information, such as names, and enables anonymised linkage across the different datasets. When data is added to the SAIL Databank, ALF assignment is completed by deterministic record linkage if NHS numbers are available and then probabilistic record linkage if deterministic record linkage is not possible. This process is detailed by Lyons et al. (2009) (see in particular Figure 1, p. 7).

In the present study, linkage of individuals was carried out across 18 datasets (with multiple schemas [or tables] per datasets). Where ALFs were available in datasets and schemas, these were used to link individuals. However, some datasets did not contain ALFs, so

corresponding ALFs were attached to individuals' records in these datasets based on other identifiers, such as pupil ID (for linking the education dataset), client ID and child ID (for linking the national child health dataset), maternal ID (for linking the maternal indicators dataset), local authority code and child code, as well as the IRN_PE code which linked to the education dataset (for linking the social care datasets).

Population

Our baseline cohort was created to both mirror the CADRE database timeframe and to include a timeframe where the most data from relevant SAIL datasets was available. The time period we identified was between 1 January 2013 and 31 March 2020. Our cohort included anyone who was aged 0–17 years in this period; all retrospective and subsequent data for these individuals was then included even if it fell outside this time period. Our final cohort consisted of 1,113,776 young people in Wales, of which 46,704 (4.19%) had social care data. This sub-sample was used for initial analysis and then for machine learning model prototyping.

Measuring risk factors for mental health problems

The risk factors of interest were identified through a three-round Delphi study (i.e. a survey of experts to reach consensus on an issue), to create a theoretical framework of risk factors for childhood mental health problems. Forty-eight experts in childhood mental health, with an average experience of ~19.5 years, were involved in developing and honing the theoretical framework, which concluded by identifying 287 risk factors. Experts included professionals from psychology, psychiatry, education, social care, public health and academia. The framework was grouped into eight domains, the first seven of which had risk factors which were potentially relevant to any young person, and the eighth domain which consisted of risk factors from the first seven domains but that may be especially salient to under-served populations. Domains one to seven are "Social and Environmental", "Behavioural", "Education and Employment", "Biomarkers", "Physical Health", "Psychological and Mental Health" and "Patterns of Service Use". Domain eight is "Factors Identified to Be Particularly Relevant to Underserved Populations". Next, we mapped the 287 identified risk factors to variables from meta-data for each of the 18 SAIL databases, creating a table of if and how each risk factor was recorded in each. We noted whether risk factors were "directly measurable", "derivable" or "measurable by proxy" within each database. We then focused on the former two types of measurability, writing code in SQL and R to measure risk factors.

Measuring childhood mental health problems

In order to measure childhood mental health problems, we used validated code lists which were shared by collaborators at the Adolescent Mental Health Data Platform (ADP). The codes were collated from published articles and code lists or were compiled in collaboration

with clinicians.[2] These code lists consisted of ICD-10 and READ codes (version 2) for: alcohol misuse; anxiety (including obsessive-compulsive disorder); bipolar; conduct disorder; depression; drug misuse; eating disorders; mood disorders (i.e. anxiety and/or depression); schizophrenia; self-harm (intentional and undetermined intent); severe mental illness (i.e. bipolar and/or schizophrenia or other psychotic conditions); and substance misuse (i.e. alcohol and/or drug misuse). We explored health datasets (WLGP, PEDW, OPDW, OPRD, NCCH, EDDS, CCDS, WRRS, SMDS, MIDS and NHSO) and the death registry (ADDE) for medical codes of diagnoses and, where available, we used them to measure the prevalence of the aforementioned mental health problems within our cohort.

Building the prototype early identification tool in social care

Our aim was to predict if a child with social care contact has a mental health problem, or not. We used a range of classification approaches. There are known challenges with building accurate classification models using electronic health records relating to data quality, such as missingness, co-linearity and accuracy of labelling (Xiao, Choi & Sun, 2018). More complex machine learning approaches can address some of these concerns, but can come with their own challenges relating to explainability and interpretability – both of which are important for models to be used in clinical settings (Meehan et al., 2022; Tonekaboni et al., 2019). Our aim was to explore a range of approaches to identify models that offer the best balance of performance, explainability and interpretability. Thus, we explored a diverse array of models ranging from statistical models, such as logistic regression, to machine learning models, such as Graph Neural Networks (GNNs). We hoped that this would allow us to gain insight into how well different models performed on multi-domain data.

Our approach was to use a logistic regression model as our baseline, and then to explore the additional benefits of other classification approaches, including XGBoost, basic Neural Network models, and Graph Neural Networks. GNNs were of particular interest to us based on recent studies illustrating their value in modelling complex relationships between related patients (Rocheteau et al., 2021) and as an approach to managing heterogeneous data with significant missingness (Malone, Garcia-Duran & Niepert, 2018), a common feature of electronic health data. GNNs allow for the simultaneous modelling of complex similarity relationships and diverse node features. Individuals are grouped together based on similar pertinent characteristics that may be predictive of their mental health status. Patients are modelled as nodes, with edges representing similarity relationships between patients, and the model allows for the sharing of information among neighbours, in order to gain a richer representation of the underlying data. GNNs tend to perform well in situations with large amounts of missing data as they can utilise the relationships inherent in the graphical structure to effectively share information across nodes during prediction (Malone, Garcia-Duran & Niepert, 2018).

---

[2] The code lists we used are available to download here: https://conceptlibrary.saildatabank.com/ADP/concepts/?collection_ids=23%2C27.

Training, testing and validation datasets

The dataset was randomly split such that 70% (32,692) of individuals fell into the training dataset, 15% (7,006) into the validation dataset, and 15% (7,006) into the testing dataset.

Variable selection, pre-processing and definition

The 287 candidate risk factors that had been identified through the Delphi study were mapped to the linked dataset, to establish which variables could be included into models. Of those risk factors that could be measured, they were excluded if they had more than 70% missing values. No variable selection methods were included in the model creation process, so all variables with less than 70% missing values were included in the final models. Moreover, all model types (logistic regression, GNN, etc.) used the same variables.

“Co-morbid physical health problems” was defined as either an individual having a formal ICD-10 diagnosis code in their clinical record or that individual having undergone an operation of some type, for example appendectomy. Diagnosis/operations codes were excluded if they had a prevalence of 2.5% or less within the cohort, to preserve patient anonymity.

To address outliers, continuous variables were standardised using sample means and standard deviations with cut-offs placed at ± 4, such that any values greater than 4 are incorporated into the model as 4 and any values less than -4 are incorporated into the model as -4. Categorical variables were included using one-hot encodings. One-hot encoding is a method that helps make categorical data amenable for machine learning methods by converting categories into a numerical vector representation.

Since the focus of machine learning methods explored here is accurate prediction rather than unbiased coefficient estimation, no assessments of collinearity were conducted.

Modelling approaches

The logistic regression model was implemented using scikit-learn. This model was utilised due to its interpretable coefficients and to set a baseline for performance. No tuneable parameters were assessed for the baseline logistic regression method and no variable selection was conducted.

XGBoost models were also applied to this task. XGBoost is a form of decision tree gradient boosting, which uses an ensemble of decision trees to make predictions. This model was included because it performs well on a variety of tasks and maintains an intermediate level of interpretability (more interpretable than deep learning methods but less interpretable than logistic regression). For the learning objective for the XGBoost model, the classifier was implemented using logistic regression for binary output classification. Hyperparameter tuning and variable selection were not conducted for this model.[3]

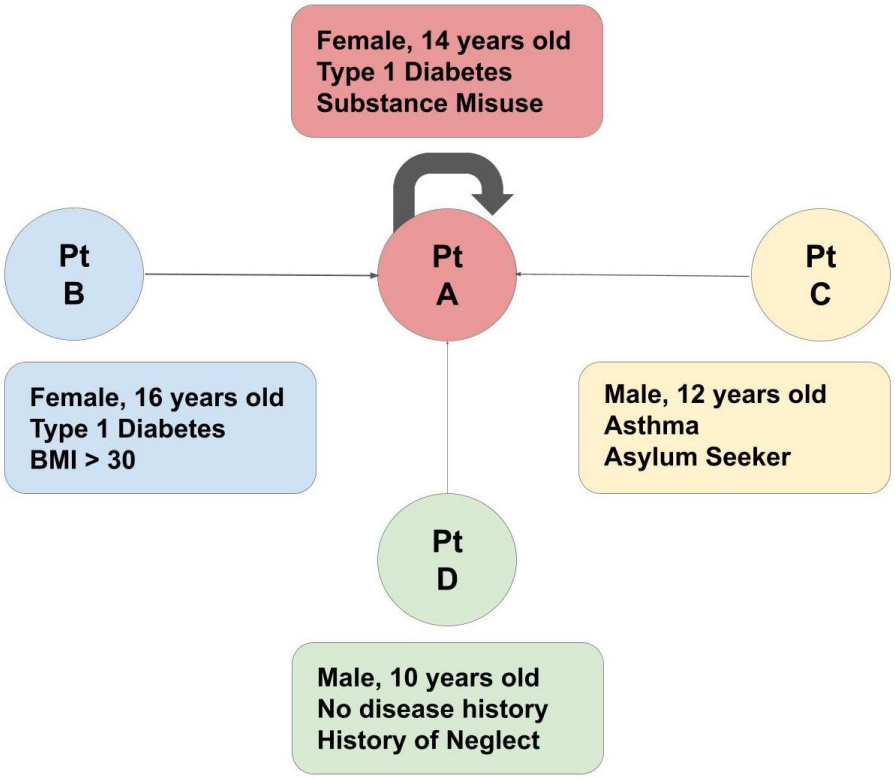
The Neural Network models employed a binary cross-entropy loss function with Rectified Linear Unit (ReLU) activations between the layers, which is a standard approach in this type

[3] For documentation discussing the implementation of XGBoost learning task parameters see: https://xgboost.readthedocs.io/en/stable/parameter.html.

of work (Sze et al., 2017). We assessed models with varying numbers of hidden layers and sizes of hidden layers in order to determine the optimal model configuration for this task. Models with more hidden layers and larger hidden layers are more expressive (i.e. can model more complex functions). Details of the hyperparameter search are discussed below.

For the GNN models, patient graphs were constructed to group together similar patients. These graphical structures consist of patients as nodes with directed edges and edge weights representing similarity relationships between patients. This general graph structure for a single example node is shown in Figure 2.1.

Figure 2.1: Sample patient graph used for GNN prediction (neighbourhood size = 3)



These graphical structures shown in Figure 2.1 consist of patients as nodes with directed edges and edge weights representing similarity relationships between patients; patients with higher degrees of similarity have a thicker line connecting them. For instance, Patient A and Patient B are more similar demographically and share a diagnosis of Type 1 Diabetes. When making predictions, GNNs take in information from a node (its own node features) and information from surrounding neighbours. The node features for the GNN are the same as the features for the other models.

We explored the three most commonly implemented layer types: Graph Attention Networks (GAT); Message Passing Neural Networks (MPNN); and GraphSAGE (SAGE) (Rocheteau et al., 2021). These layer types were utilised because they differ in their expressivity. Expressivity refers to the complexity of functions that can be modelled using a given machine learning approach, where more expressive models can model increasingly complex functions. Hence, expressive models can perform well in situations where there is a

complicated non-linear relationship between input variables and output variables; however, more expressive models have a tendency to overfit noise within the training data. To find GNN models that optimally balance this trade-off between expressivity and overfitting, layers differing in expressivity were assessed, with MPNN allowing for the greatest modelling flexibility and the SAGE and GAT models providing less expressivity.

Hyperparameter search

Hyperparameter searches were performed to identify the optimal hyperparameters amenable to this particular dataset. For the logistic regression and XGBoost models, no tuneable hyperparameters were assessed. For the neural network models, hyperparameter searches over the learning rate, size of hidden layer (corresponding to model complexity) and number of hidden layers were conducted using the validation dataset. A hyperparameter search for the GNNs over the validation set was conducted to identify the optimal learning rate, size of hidden layer and neighbourhood size for each layer type (Graph Attention Networks; GraphSAGE; Message Passing Neural Network).

Comparison of model performance

We compared the performance of the XGBoost, Neural Network and GNN to the baseline logistic regression model using AUROC. When assessing the performance of these approaches, we favoured false positives in order to reduce the likelihood of missing a genuine case of a mental health diagnosis. This choice will probably have led to more identified cases than the true ones, but we consider this to be preferable compared to missing cases, and further investigation including alternative analytical approaches and sensitivity analyses will identify these cases.

3. Findings

Describing our cohort

Our baseline cohort of young people, based on the Welsh Demographic Service Dataset (WDSD), consisted of 1,113,776 people, with 50.74% male, 49.26% female and <0.01% unknown. Information on individuals identifying as non-binary was unavailable. A sub-sample of this baseline population was used for prototype early identification model development; this sub-sample comprised individuals with social care records (46,704 people, of which 53.91% were male and 46.09% were female). Demographic details of biological sex, ethnicity (where available) and age are presented below in Table 3.1. Age was calculated for all individuals as of 1 January 2022. Of note, given we had all data relating to an individual if they fell within our cohort definition, this means we also had data from some individuals aged over 17 years old too. Ethnicity data was not available in WDSD for the cohort. However, ethnicity data was available for the sub-sample of young people who had access to social care services included in our prototype early identification models. Of note, though the overall cohort comprises 1.1 million young people in Wales, sample sizes differ quite substantially between datasets (as shown below in Table 3.2).

Table 3.1: Characteristics breakdown of our baseline cohort compared with the sub-sample who had social care contact (i.e. sub-sample used for prototype early identification model)

Characteristic	Characteristic category	Number of individuals in baseline population (N (%))	Number of individuals in social care sub-sample (n (%))
Biological Sex	Male	565,086 (50.74)	25,179 (53.91)
	Female	548,673 (49.26)	21,525 (46.09)
	Unknown	17 (<0.01)	0 (0)
Total		**1,113,776 (100)**	**46,704 (100)**
Ethnicity	Asian	[Unable to measure]	855 (1.83)
	Black	[Unable to measure]	542 (1.16)
	Mixed	[Unable to measure]	1,317 (2.82)

	Other	[Unable to measure]	3,122 (6.68)
	White	[Unable to measure]	40,868 (87.50)
Total		**1,113,776 (100)**	**46,704 (100)**
Age	0–2	81,995 (7.36)	0 (0.00)
	3–5	106,447 (9.56)	638 (1.37)
	6–8	116,062 (10.42)	3,739 (8.01)
	9–11	126,524 (11.36)	6,404 (13.71)
	12–14	129,294 (11.61)	8,194 (17.54)
	15–17	124,504 (11.18)	8,814 (18.87)
	18–20	135,676 (12.18)	8,053 (17.24)
	21–23	164,010 (14.73)	6,873 (14.72)
	24–27	129,255 (11.61)	3,989 (8.54)
Total		**1,113,776 (100)**	**46,704 (100)**

Coverage of datasets

We characterised each of the 18 datasets and explored the proportion of individuals from the baseline cohort who were present in each dataset (presented in Table 3.2). This information can be interpreted in combination with risk factor measurement in order to understand how valuable each dataset is for developing prediction models. For example, a risk factor may be measurable in multiple datasets and this coverage information can be used to select which dataset(s) should be used for measuring the risk factor.

Table 3.2: Table to show the number and percentage of individuals from our cohort who are present in each dataset, presented by database type

Dataset type	Database name	Number of individuals	% of cohort
Demographics	Welsh Demographic Service Dataset (WDSD)	1,113,776	100
	Annual District Birth Extract (ADBE)	800,714	71.89
	Annual District Death Extract (ADDE)	2,400	0.22
Education	Pre-16 Education Attainment (EDUW)	766,250	68.80
Social Care	Child In Need – Wales (CINW)	35,481	3.19
	Children Receiving Care and Support (CRCS)	26,653	2.39
	Looked After Children – Wales (LACW)	7,363	0.66
Healthcare (General Practice)	GP Primary Care (WLGP)	944,933–1,095,022*	84.84–98.32*
Acute Healthcare	Critical Care Dataset (CCDS)	2,043	0.18
	Emergency Department Dataset (EDDS)	710,752	63.81
Community Healthcare	National Community Child Health (NCCH)	290,391–999,375*	26.07–89.73*
	NHS 111 Call Data (NHSO)	165,790	14.89
	NHS Hospital Outpatients (OPDW)	694,418	62.35

	Outpatient Referral Data (OPRD)	629,193	56.49
	Patient Episode Database for Wales (PEDW)	629,904	56.56
	Substance Misuse Dataset (SMDS)	4,604–11,545*	0.41–1.04*
	Wales Results Reporting Service (WRRS)	876,039–877,005*	78.65–78.74*

*Some of the datasets are made of sub-datasets (called schemas) so contain a range of individuals depending on the specific schema used within the dataset.

Research Question 1: What is the best method of measuring mental health problems and risk factors for young people's mental health problems in linked administrative datasets?

In the following sections, we present the measurement of risk factors associated with mental health problems, followed by the measurement of mental health problems in our cohort.

Measurement of risk factors associated with mental health problems

We mapped the 287 risk factors identified through the Delphi study to variables using meta-data for each of the 18 SAIL databases, creating a table of if and how each risk factor was recorded in each of the 18 databases. We found that there was a spectrum of measurability for the risk factors. This spectrum can be considered in terms of three categories: whether risk factors were 'directly measurable', 'derivable' or 'measurable by proxy' within each database (see Table 3.3 for definitions and examples of each type).

Table 3.3: Table to describe how we defined measurability of risk factors

Name of measurement type	Definition	Examples
'Directly measurable'	'Directly measurable' risk factors were defined as those which could be measured using the variable alone, without manipulation of the variable or taking into account information from other sources (including other variables). This means a variable exists in a dataset	'*Gender*' which is captured by the variables 'GNDR_CD', 'PAT_SEX_CD', 'GENDER' and 'SEX'. '*Participation in the Free-Lunch Programme*' which is captured by the variables 'FSM' and 'FSMELIGIBLE'.

	or datasets which directly describes this risk factor.	'*Child protection record*' which is captured by the variable 'child_protection_register'.
'Derivable'	'Derivable' risk factors were defined as those which could be measured by manipulation of information from one or more variables, or taking into account information from other sources.	'*Body Mass Index (BMI)*' can be calculated by using the variables height and weight. '*Three or more presentations to emergency services within a year*' can be measured by summing distinct events from emergency service datasets. '*Anaemia*' can be measured by using the variable containing diagnostic codes in health datasets and identifying codes consistent with a diagnosis of anaemia (e.g. read codes, ICD-10 codes). (Of note, measuring risk factors such as this is likely to require discussion with a clinician(s) to identify relevant diagnostic codes, as opposed to work which can be carried out by the research team alone.)
'Measurable by proxy'	'Measurable by proxy' risk factors are those which can be estimated or inferred using a variable or a combination of variables. These measurements are an approximation or indicator of the risk factor in question, rather than being a direct or derived measure of it. There will be a greater degree of subjectivity in these measurements.	'*Low socioeconomic status family*' could be inferred using proxies such as employment status of parents/caregivers, lower super output area (LSOA) and index of multiple deprivation (IMD), and variables pertaining to eligibility for free school lunches. '*Chronic (long lasting) infection (e.g. Lyme disease, periodontal disease)*' could be inferred through data on prescriptions for medication to fight infections. (Of note,

		it could be also be derived as described above using diagnostic codes). '*Household mental illness*' could be inferred from the antenatal health check datasets using a variable which indicates that the mother has a health care plan for mental health problems. Of note, this would only relate to the mother, not the whole 'household' and would not identify mothers who previously had care plans but no longer do, and it would also not cover mothers who go on to develop mental health problems later on in the child's life (i.e. it only describes the child's circumstances at birth).

In the interest of time, we focused our efforts on measuring directly measurable and derivable risk factors. We reviewed each database and recorded if it contained a relevant risk factor, illustrated in Table A1 and B1 (see Appendix A and B). In Table A1, we show the number and description of the directly measurable and derivable risk factors that can be measured for each of the 18 SAIL databases. With the aim of assessing the value of multiagency data for predictive modelling in this context, we grouped the databases into types (demographics, education, social care, primary care, and specialist or acute care), to illustrate the extent of coverage of variables in each database type. Some risk factors were measurable in multiple datasets; however, it is important to still use all of the datasets because not every individual young person appears in every dataset (as highlighted in Table 3.2 above). Table B1 aims to highlight the extent to which we are able to measure risk factors identified as important to underserved populations.

Of 287 risk factors of interest from domain one to seven of the theoretical framework, 101 (35.19%) were measurable; of these, 48 (16.72%) were directly measurable and 53 (18.47%) were derivable (see Table A1). Of 101 risk factors of interest relating to underserved populations (i.e. domain 8 of the theoretical framework), 37 (36.63%) were measurable; this is broken down into 26 (25.74%) which were directly measurable and 11 (10.89%) which were derivable (see Table B1). We describe these as qualitative tables because this risk factor mapping exercise led us to conclude which factors we consider to be measurable based on SAIL meta-data. Extracting these risk factors using code will allow us to conclude to what extent they are actually measurable.

We illustrated that the risk factors considered to be important to predict childhood mental health problems cannot be predominantly found in one database, or database type.

Demographic data was measured across two core databases, with important information such as urbanicity, birth weight and employment of primary caregivers found in the Annual District Birth Extract, and deprivation, sex and month of birth from the Welsh Demographic Service Dataset. Eight variables were measurable in the education dataset. They provided important information on ethnicity, as well as indicators of deprivation such as participation in the free-lunch programme. Information about educational attainment and behavioural proxies such as out-of-school discipline and exclusions were available. Measures of disability were available through markers for "special educational needs". Social care datasets provided important information about adverse childhood experiences, such as domestic violence, abuse or neglect, household substance misuse and involvement in the criminal justice system. The degree of social care contact provided an indicator of need (e.g. looked after child status). Primary care data was of particular use for measuring co-morbid physical health conditions. Specialist and acute healthcare datasets included A&E and 111 data, postnatal specialist care, paediatrics, outpatients, substance misuse and pathology results. These provided important validation for GP data on physical health co-morbidity, as well as a range of pre- and post-natal indicators such as Apgar scores, breastfeeding, maternal smoking, birth weights, prematurity and BMI. A range of biomarkers were available such as levels of folate, iron and vitamin D.

Measurement of childhood mental health problems

Measurement of the prevalence of childhood mental health problems in our cohort is presented in Table C1 (see Appendix C), using data from our reference health datasets. We used the Welsh GP dataset (WLGP) as our baseline, finding that 14.85% of our cohort had at least one mental or psychological health condition of interest, with mood disorders being most common (12.96%) and severe mental illness (SMI) such as schizophrenia and bipolar being least common (0.11%). For comparison, we measured the rates of mental health disorder in the Patient Episode Dataset for Wales (PEDW), which includes data on patients in contact with acute and outpatient services, and found that the prevalence of any mental health condition of interest was 4.78%, with mood disorders still being most common (2.73%) and SMI being least common (0.13%).

Research Question 5: What are the best methods for building predictive risk models and early identification tools for young people's mental health problems for use in social care settings?

In total, there were 46,704 unique individuals with social care data who formed the cohort explored in this sub-analysis. Categorical and continuous variables included in the models are shown in Tables 3.4 and 3.5 respectively.

Table 3.4: Categorical variables included in the model

Domain in Delphi framework	SAIL variable name	SAIL dataset
Domain 1: Social and Environmental	Asylum Seeker Status	CINW/CRCS

	Breastfeed Status (Birth)	MIDS
	Breastfeed Status (8 weeks)	MIDS
	Category of Need	CINW/CRCS
	Looked After Child Status	CINW/CRCS
	Maternal Smoking	NCCH
	Parenting Capacity (Domestic Abuse)	CINW/CRCS
	Parenting Capacity (Mental Health)	CINW/CRCS
	Parenting Capacity (Substance Misuse)	CINW/CRCS
	Urban/Rural Status	ADBE
	Youth Offending Status	CINW/CRCS
Domain 2: Behavioural	Substance Misuse	CINW/CRCS
Domain 3: Education and Employment	Free School Meal Status	EDUW
	Parenting Capacity (Learning Disabilities)	CINW/CRCS
	School Exclusion Category	EDUW
Domain 4: Biomarkers	Gender	WDSD
Domain 5: Physical Health	Dental Check Status	CINW/CRCS
	Disability (Mobility)	CINW/CRCS
	Disability (None)	CINW/CRCS
	Labour Onset	NCCH
	Parenting Capacity (Physical Health)	CINW/CRCS
Domain 6: Psychological and Mental Health	Autistic Spectrum Disorder Status	CINW/CRCS
	Disability (Memory)	CINW/CRCS
	Disability (None)	CINW/CRCS
	Disability (Sensory)	CINW/CRCS
Domain 7: Patterns of Service Use	Child Protection Register Status	CINW/CRCS
	Health Surveillance Checks Status	CINW/CRCS
	Immunisation Status	CINW/CRCS

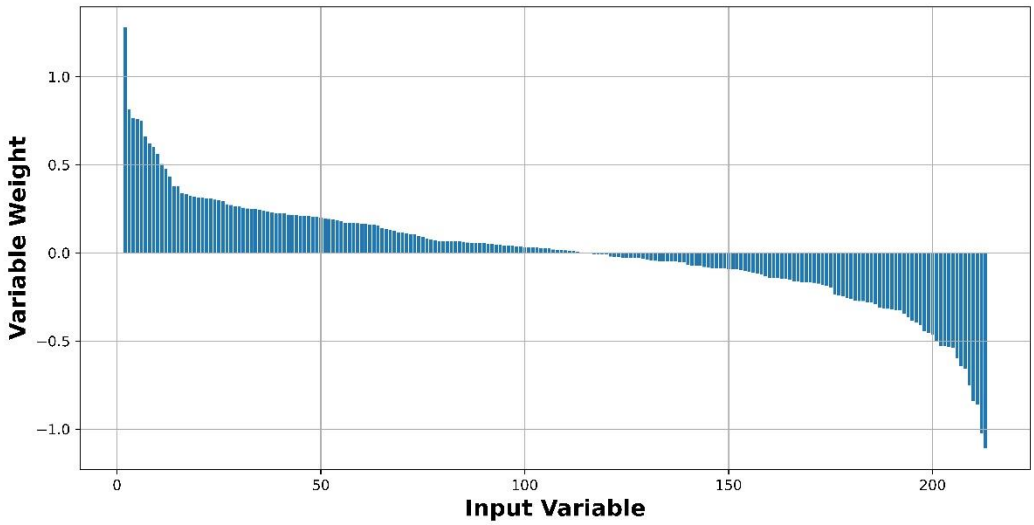
Table 3.5: Continuous variables included in the model

Domain in Delphi framework	SAIL variable name	SAIL dataset
Domain 1: Social and Environmental	Age	WDSD
	Welsh Index of Multiple Deprivation	WDSD
Domain 2: Behavioural	_	_
Domain 3: Education and Employment	_	_
Domain 4: Biomarkers	Birth Weight	NCCH
Domain 5: Physical Health	Apgar 1-Minute	NCCH
	Apgar 5-Minute	NCCH
	Gestation Age	NCCH
Domain 6: Psychological and Mental Health	_	_
Domain 7: Patterns of Service Use	_	_

Interpretability of logistic regression model

The interpretability of the logistic regression model was assessed to gain insight into model performance and the variables important for model prediction. Model weights for each of the different risk factors are shown in Figure 3.1 below, with the risk factors ordered from most predictive of positive mental health status to most predictive of negative mental health status. Odds ratios for each individual risk factor are presented in Table D1 (see Appendix D).

Figure 3.1: Variable importance analysis: Ordered list of raw weights provided by logistic regression model for each variable*



* Of note, there are more variables in Figure 3.1 than outlined above because some risk factors had multiple sub-categories (e.g. the risk factor “Ethnicity” included Irish, Indian, etc.).

No single variable dominates the model which indicates that a variety of features play a role in the prediction of child mental health. The risk factors that are most predictive of a mental health diagnosis were: substance misuse (largest coefficient of 1.28), having autism (coefficient of .81) and being Indian (coefficient .77). Here, the coefficient of 1.28 relating to substance misuse converts to an odds ratio of 3.6. This indicates that, for someone who misuses substances, the odds of having a mental health diagnosis are 3.6 times higher than the odds for someone who does not misuse substances. Interestingly, all of these risk factors relate to characteristics specific to the individual, rather than characteristics relating to an individual’s family situation, for example physical ill health of parents or living in a household that experiences domestic abuse. This may be a reflection of the quality of the data relating to individuals compared to families, or may reflect the relative importance of the risk factor depending on the proximity of its effect. The former explanation could be confirmed or disconfirmed by linking family data within SAIL (which is possible using residential ALFs) and measuring family-level risk factors directly. At present, some of these risk factors are measurable or inferable with the child’s data alone, but this would be enhanced by measuring directly from family members’ records. For example, parental ill health can be inferred from social care flags when parenting capacity is impaired by ill health, but this will only capture more extreme instances of this risk factor. Conversely, by linking the parents’ data, parental ill health could be measured more directly through their own medical interactions (e.g. appointments, prescriptions and diagnoses). Another example would be maternal smoking status, which could currently be measured from the smoking status at the time of the child health assessments in the MIDS or NCCH, but it could also be directly measured if the mother’s data was linked and contained her smoking status. Alternatively, the latter explanation of proximity effects would be consistent with

Bronfenbrenner's person-process-context-time theory, which hypothesises that risk factors have greatest impact when most proximally affecting an individual (Bronfenbrenner, 1995). Further work is required to understand these alternative explanations better, i.e. if we measured risk factors as well as we feasibly could using data linked within families and between databases, do individual-level risk factors still appear most significant or is this an artefactual finding relating to the varied ability to measure child-level vs family-level risk factors via the child's data?

In the social care datasets, there are "mental health status flags"; this positive mental health status corresponds to individuals diagnosed with mental health problems by a medical practitioner or individuals on a waiting list for such services. These diagnoses include mental health problems such as depression, eating disorders and self-harm, but they do not include learning disabilities or substance misuse if they are not accompanied by other mental health issues. Individuals were considered as having a positive mental health diagnosis if they contained a positive flag in either the CINW or CRCS dataset at any point. Utilising the data provided from the social care datasets, we measured the mental health status of the cohort (results shown in Table 3.6).

Table 3.6: Mental health status of social care sub-sample

Mental health status	Number of individuals (N)	Proportion of cohort (%)
Positive	6,706	14.36
Negative	39,998	85.64
Total	**46,704**	**100**

The dataset was imbalanced since the vast majority of individuals within the dataset (85.64%) did not have a diagnosed mental health problem. To address this imbalance, the weight of the training samples were adjusted to obtain a classifier with real-world utility. The equations for this are shown below:

$$Positive\ Weight = \frac{Number\ of\ Samples}{2*(Number\ of\ MH+\ Individuals)}$$

$$Negative\ Weight = \frac{Number\ of\ Samples}{2*(Number\ of\ MH-\ Individuals)}$$

In these equations, MH+ refers to children with a mental health problem, while MH- refers to children without a known mental health problem. The next stage of the analysis was to test the added value of a range of alternative models.

Hyperparameter search

After conducting hyperparameter searches using the training data, the best-performing neural network model had four layers and 60 nodes per layer, and it used a binary cross-

entropy loss function with Rectified Linear Unit (ReLU) activations between the layers. The best-performing SAGE model had a learning rate of .01, four neighbours per node, and size of hidden layer of 128. The best-performing GAT model had a learning rate of .005, two neighbours per node, and size of hidden layer of 16. Finally, the best-performing MPNN model had a learning rate of .01, two neighbours per node, and size of hidden layer of 32. The logistic regression and XGBoost models employed contained no tuneable hyperparameters so no hyperparameter search was conducted for them.

Summary of model performance

Performance of the various models on the test set can be seen in Table 3.7. Area Under the Receiver Operating Characteristics Curve (AUROC) was utilised as the primary measure due to the class imbalance within the dataset. In situations with extreme class imbalance such as this task, accuracy is not a suitable metric as models can trivially achieve high accuracy by always predicting the majority class. Performance varies between runs for some models due to different model initialisations, so performance was averaged over 15 different runs and the 95% CI for mean performance for error margins is reported.

Table 3.7: Model performance

Model	AUROC Test
Logistic Regression	.803
XGBoost	.800
Neural Network	.800 ± 1.1e-3
Graph Attention Network (GAT)*	.783 ± 9.1e-4
Message Passing Neural Network (MPNN)*	.807 ± 6.9e-4
GraphSAGE (SAGE)*	.815 ± 7.0e-4

Note: Graph Neural Network (GNN) models indicated above with an * asterisk.

All models demonstrate similar levels of performance, with the GraphSAGE model achieving the highest level of performance, although this was marginal. The neural network, logistic regression model, and XGBoost model all achieve similar performance. Given that the goal is to use similar models in clinical practice, it is important to assess the relative interpretability of the different models. Logistic regression models are considered the most interpretable models assessed as each input variable is directly given a weight, which

means it is possible for clinicians to understand the relative importance of various predictors to the model. This may provide a target for intervention of modification of risk. In contrast, the XGBoost and neural network models are less interpretable than the logistic regression model but can model more complex relationships which may lead to improved performance. In this particular task, the additional modelling flexibility did not translate to improved performance, indicating that simpler, more interpretable models may be preferable.

Research Question 6: Can findings and methods be replicated across databases (i.e. translated to CADRE)?

The work carried out within SAIL has important learning points and implications for the CADRE database we are building in Cambridgeshire and Peterborough. Furthermore, work carried out in SAIL (facilitated by WWCSC funding) has allowed us to secure subsequent funding to enable the development of a network of Trusted Research Environments (TRE), starting with the Cambridgeshire and Peterborough based database, CADRE and now extending into the development of a federated network of TREs including Birmingham and Essex. This network of TREs will be important for next steps, as it provides a means to externally validate models built in Cambridgeshire and Peterborough, testing their generalisability to different populations. Below we summarise this learning, both in terms of the development of TREs and early identification model prototypes. This learning can then be used to enhance future work.

Implications and learning points for the TREs in development:

- Getting started with work in SAIL was challenging and there was a steep learning curve with understanding the databank and databases. This could have been alleviated with a clear step-by-step new starter guide, including set-up instructions. Furthermore, support information that did exist was protected behind the secure gateway meaning that we could not make the most efficient use of time whilst awaiting IGRP approval. It could be helpful to include new starter guides and user information in an accessible place (instead of on the intranet or internal gateway) so that new database users can become more familiar with the database while awaiting approvals, thus allowing them to get started quickly and effectively if approvals are granted. This would also ensure that potential users are in the best place to make an informed decision about whether the database matches their needs and competencies when deciding whether to apply to access the databank. Naturally these user guides would not contain any secure or confidential information but would simply be technical guides.
- The SAIL application process gave us a model of one possible method of researchers accessing our database. For example, we learnt about the pros and cons of SAIL's way of managing the project-scoping phase, application form, internal and external review panel, technical support, computing power, etc. This insight can help us to translate the best bits of SAIL's processes and improve on those which were a barrier for us.
- We learnt how important it is to leave significant amounts of time for data cleaning and processing when working with such vast databases. This will help us set realistic time frames for prospective projects, as well as help us to understand how some of the data-processing pipeline could be semi-automated.

- The risk factor mapping exercise highlighted the importance of having clear, accurate and detailed meta-data for each variable. We frequently had to seek clarification on the element-level data of a variable in order to understand what a variable was and how it could be used.
- The expertise and inside knowledge of colleagues working frequently with SAIL databases was invaluable, including for linking data, sense checking information, and interpreting results. It will be important to have a mechanism for sharing such expertise in our work, for example through on-hand technical support or user forums.
- Through this work, we formed invaluable collaborations with related researchers, specifically Professor Ann John's team in the ADP. These collaborations will allow us to combine efforts and produce more efficient and high-quality research on young people's mental health problems.
- It was clear that it is important to have robust computing infrastructure with sufficient processing power to run complex models. Computer crashes occurred frequently, especially when working with large datasets like the WLGP, and this limited our rate of progress. For research projects which are likely to run on short time frames, including short funding cycles, this is an important issue to avoid in our database development. In addition to computing infrastructure, having research support from people with knowledge of the process of machine learning development would have been invaluable, especially in relation to carrying this out in SAIL and with the expertise to comment on the trade-offs of this.
- Work in SAIL further highlighted the importance of demographic data, including ethnicity data, for developing fair and equitable models. The limited demographic information and lack of ethnicity data in most of the SAIL databases was a limiting factor and something we wish to address in the databases we build.
- The exercise of mapping risk factors to SAIL meta-data led us to conclude that there is a spectrum of measurability (as defined in Table 3.3). In the other TREs in development, we would repeat this mapping exercise, with an initial focus on "directly measurable" and "derivable" risk factors. Though the exact variable names will inevitably change between databases, this mapping exercise should hopefully help us more quickly pinpoint where equivalent risk factors are in each TRE, expediting model development. Furthermore, mapping risk factors in SAIL helped us to consider how to operationalise the measurement of each risk factor (i.e. how to decide if an individual is classed as having or not having a particular risk factor). More specifically, this was because we were able to learn about how the variables (i.e. risk factors) tended to be recorded in routinely collected data. This work will form the basis of a data dictionary and data model for child mental health prediction, which will be developed for CADRE.
- The absence of time variables was a limiting factor when developing early identification models in SAIL and longitudinal analysis was hindered. We plan to address this in our database build by ensuring all events are time-stamped, allowing for more dynamic interactions to be explored and temporal information to be taken into account.

Implications for prototype early identification models:

- This exercise of prototype development was our first attempt at developing early identification models for children in social care. It gives a foundational performance

which we can build on and refine, by incorporating additional predictive risk factors. Not all of the risk factors identified to be important to predicting mental health problems were measurable, and developing methods to measure the remaining risk factors will provide us with additional predictors for inclusion in models.

- With machine learning methods, it is vital to externally validate models using different databases/populations, to ensure the models can generalise beyond the database and population from which they were originally developed. Work in SAIL provides a prototype for further development, and external validation. It will be important to validate in diverse populations.
- From this work, we can better understand what additional value each single dataset brings such that we could decide which datasets are a minimum requirement, which are desirable, and which may be superfluous when trying to build accurate risk prediction models.

Research Question 2: What is the prevalence and distribution of mental health-associated problems and their risk factors? How do patterns of mental health-associated problems vary between social care, health and educational settings? How do they vary across Wales, UK?

Research Question 3: What is the unrecognised mental health need in social care settings?

Research Question 4: What are the relationships between risk factors and mental health problems?

Due to the unexpected complexity and time requirement to carry out data linkage, cleaning and pre-processing, we were unable to address Research Questions 2–4 within the scope of this funding cycle. We plan to carry out prevalence visualisations in the next phase of the project now that we have created a linked database and have a greater understanding of the datasets and their variables.

4. Limitations

Technical/computing limitations

Throughout our analysis, computer memory issues have hampered measurement, leading to frequent system crashes and requiring coding workarounds. This is because there are a vast number of rows of data per individual in datasets such as WLGP and when combining this dataset with others, the volume of data exceeds the system settings. However, we did not alter or simplify our analysis plan; in order to address this, we have been working with the SAIL support team to overcome the limitations, including successfully securing additional funds to pay for enhanced computing power.

Dataset-level limitations

There are certain limitations associated with the datasets included in this study. With regards to the GP dataset, SAIL currently has data agreements with >70% of Welsh GP surgeries so this data covers the majority of the population, but a non-trivial number of practices are missing from the databank at this time. If the 70%+ GP surgeries included in the databank are representative of the general population, and characteristically similar to the CADRE population, then this is not necessarily a problem in the present study. However, it is possible that the Welsh GP surgeries who do not have data agreements with SAIL systematically differ from those with agreements. To address this, SAIL are continually developing collaborations with providers to increase their data coverage. External validation would be an important part of any validation approach and would help to mitigate the risks associated with non-representative datasets.

With regards to the available social care datasets, the “Child In Need – Wales” and “Children Receiving Care and Support” dataset are census returns. As noted by Lee et al. (2022), who analysed this data too:

> *“[As census returns, they are] only able to offer an annual snapshot into the circumstances of eligible children. Children might be missed from a return if they join or leave the cohort of eligible children outside the return dates. It is also not possible to capture every state that a child might experience throughout the year. For example, a child might be recorded as not having a child protection plan in place, but they could join the child protection register in mid-April of the same year, and the return for that year will have no record of this if they come off the register again before the next census period.”*

This lack of granular social care data likely limits the predictive accuracy of models developed. In spite of these limitations, SAIL is an invaluable resource, providing real-world, administrative data, which can be used to understand the experiences and needs of young people, with important implications for health and social care. Increasing the frequency of data returns or including longitudinal data may help to mitigate these risks.

Data-level limitations

Data quality and accuracy is also a limitation in this project. As with all studies, the results and models are only as accurate as the data they are built on. Data accuracy in this project is affected by a couple of different things: diagnostic assignment accuracy and diagnostic recording practices. First, inaccuracies in diagnostic coding will affect our results; for example, different diagnosing clinicians will diagnose mental health conditions differently and the inter-rater reliability between diagnosing clinicians in mental health has previously been demonstrated to be quite poor at times (see for example Davis, Sudlow & Hotopf, 2016; Matuszak & Piasecki, 2012; Nicholls, Langan & Benchimol, 2017; Regier et al., 2013). Our results assume that the mental health diagnostic labels are accurate for each individual, but this will never be entirely true. The accurate attribution of diagnostic codes or labels will vary within diagnosing clinician, between diagnosing clinicians, between healthcare providers, and over time, to name but a few variations. Second, inaccuracies will inevitably be present from the recording of diagnostic codes (see for example Davis, Sudlow & Hotopf, 2016; Nicholls, Langan & Benchimol, 2017); as with the application of diagnostic codes, the recording of such codes will also vary within the same clinician at different times, between clinicians, between providers, and over time (for example, following the implementation of new policies in an organisation or increased focus on certain conditions). The accuracy of our results is inevitably limited by diagnostic practices and recording practices, and our results are predicated on the assumption that diagnoses are both accurate and recorded, though this will not be true all of the time. This limitation is true of any research using labels of any kind; however, it is especially true for real-world healthcare data which is collected for clinical use, not research use, and is inevitably “messy” data. However, the benefits of using such data outweigh the drawbacks; this data is a closer reflection of reality than research data would be and is also more comparable to where any early identification models would be implemented and to the data these models would be utilising. This means that routinely collected data is still the most appropriate choice of data source for this project.

An additional data-level limitation relates to the lack of temporal data. SAIL datasets contain very little information relating to the timing of events, limiting our ability to model longitudinal change or include the dynamic nature of risk factors when creating the prototype risk prediction tool. Health events, including diagnoses, were timestamped; however, the majority of the other variables and risk factors were not, which limited our analysis. Where available, we incorporated time information into models developed within SAIL. As with the measurement of mental health prevalences, our results for the measurement of risk factors are predicated on the assumption that risk factors are accurately recorded. Furthermore, the lack of time data means that we were unable to apply Bronfenbrenner’s model properly. If we could include more temporal information for risk factors, we could be better able to predict mental health problems with increased accuracy because timing and repeated exposure to risk factors could be included in the models. In CADRE, we plan to include timestamps on all events to ensure longitudinal analyses can be carried out robustly and this was an important learning point from analysis in SAIL. Moreover, CADRE will contain free-text in the future, not just structured data as SAIL does. This will allow us to employ natural language processing (NLP) methods to extract information on risk factors and mental health problems from free-text notes, which is likely to improve prediction model accuracy. Through the Delphi study, we have a prioritised list of risk factors to target NLP algorithm

development efforts. NLP methods may help us to measure some of the remaining 65% of risk factors from the Delphi study which could not be directly measured or derived thus far.

Risk prediction prototype limitations

The poor availability of demographic information, such as ethnicity data, for the whole cohort in this study was a particular limitation for model development. It is important to know whether prediction models perform well for people from different backgrounds, especially to ensure that underserved minority groups are not further underserved by models. The social care datasets did contain ethnicity data, which was incorporated into the prediction model; however, ethnicity data was not available for the rest of our cohort who did not have a social care record. This information would be important when developing models in other care settings, such as acute healthcare. Incorporation of data on demographics such as ethnicity and socioeconomic status, and subsequent evaluation of model performance by sub-groups, would be crucial steps before any prediction models could be deployed in public-facing or clinical settings. It is also important to note that great caution should be utilised when incorporating in ethnicity data into models which influence access to care to avoid exacerbating existing disparities.

Progress is being made in this area, however. The SAIL team are starting to develop a computing package for providing categorised ethnicity data for individuals; in the future, projects accessing SAIL would be able to apply for access to this package once it has been developed. Projects can also currently apply to access and link the 2011 census records to gain information on the ethnicity of some individuals; however, this was a lengthy process and thus we opted not to pursue this route given the time-sensitive and resource-intensive nature of this pilot project, but incorporation of more ethnicity data in future would be an important step. The SAIL team are also in discussion with the Office for National Statistics to acquire a more current census dataset (i.e. the 2021 census).

Another important consideration with regards to risk prediction model limitations is whether the model accurately predicts mental health problems rather than mental health service access. This was an important distinction when considering what outcome variables to include in the model development. Diagnostic codes alone as an outcome would be insufficient because some young people will have emerging mental health problems but remain undiagnosed, for example due to lack of access to secondary services. For our outcome measures, we used code lists for mental health problems which were developed and validated by Professor John and colleagues (Economou et al., 2012; John et al., 2021, 2018, 2016; McGregor et al., 2010; Rees et al., 2022; Wood et al., 2019). These code lists allowed us to more robustly measure mental health problems, making use of considerable prior research to validate these code lists so that they incorporated diagnostic codes, as well as proxies for mental health problems (such as prescriptions). Though not completely water-tight, the use of proxies like prescription data for mental health problems, not just mental health diagnoses themselves, was crucial so that we could get closer to developing a model which predicts actual mental health problems, rather than just predicting access to mental health services (i.e. access to someone who could diagnose a mental health problem). This is important so that any developed risk prediction models do not just predict the most severe instances of mental health problems (i.e. those which are likely to get diagnosed in mental

health services) but also more mild and moderate levels of problems. Furthermore, it is important to include proxies in order to mitigate the risk of widening health inequalities (i.e. avoiding a model which favours identifying those who already have access to secondary care, and disadvantages those who do not access secondary services but have need of them).

Though the early identification models show promise and performed well, they are not yet accurate enough to be used in clinical settings. However, this exercise has formed a valuable prototype exercise, helping us to understand where predictive risk factors might be and how models might perform for subgroups such as those with social care data. This information is a crucial foundational step to building accurate, robust, fair and generalisable early identification tools which are fit for clinical use.

5. Discussion

Through analysis in the SAIL Databank, we were able to explore childhood mental health problems and their associated risk factors, culminating in the development of early prototype risk prediction models which offer some promise, and provide foundational models for further refinement. Previous work to develop prediction models in children's social care have been unable to reach the performance needed for clinical use (Clayton et al., 2020). We hypothesised that this was in part due to the use of single-organisation data, pertaining to social care alone, as well as insufficient sample size. Mental ill health is increasingly being understood as a bio–psycho–social issue, with various associated risk factors. As such, we hypothesised that building accurate models to identify mental health problems would require access to large, representative datasets of multi-agency and multi-domain data reflecting a broad range of bio–psycho–social factors. Accessing and linking different datasets within the SAIL Databank afforded us the opportunity to more holistically explore childhood mental health problems and associated risk factors.

Our analysis illustrates the value of multi-agency, administrative data. Single data sources, such as children's social care data, provide some valuable information about risk factors for mental health problems, but when this is combined with other data sources such as those from healthcare (like patient episode data and GP data), we can more comprehensively explore mental health problems and needs in populations of young people, and begin developing predictive models for these groups. By creating a linked dataset, we have laid the foundation to compare the prevalence of mental health problems and associated risk factors in different single data sources with the prevalence for the population when this information is explored in a linked, multi-agency way.

Data linkage projects on mental health trajectories often omit social care and other local authority data (such as council data), instead tending to favour data linkage between different health data types (such as mental health and acute health), or between education (such as the National Pupil Database) and health data (such as Hospital Episode Statistics) (see for example Downs et al., 2017; Grath-Lone et al., 2021; John et al., 2021). However, our analysis has highlighted what an asset social care data is when combined with these other data sources. Social care data provides rich data on a small but significant subgroup of young people with additional care needs and potentially different risk factor profiles. This data was incredibly valuable when building a prototype early identification tool because of the breadth of risk factors included within it, as well as its low data missingness, in contrast to other data sources. This analysis showcases the worth of including social care data in data linkage projects such as CADRE. Although linkage of datasets without unique identifiers such as NHS numbers is complicated, much is gained from efforts in this area.

Other important efforts have been made to link health, education and social care data (Downs et al., 2019). However, this study utilised a clinical cohort, with health records only being available for young people who already had a CAMHS record (i.e. a mental health record) in South London and Maudsley (SLaM) health trust, and social care data was extracted from the CAMHS records rather than actually incorporating social care datasets. Crucially, when trying to identify or predict mental health problems earlier, the data sources need to be as near to whole-population as possible so that pre-clinical and non-clinical cases

rather than clinical cases can be detected. By utilising the SAIL Databank, we were able to capture closer to the whole population, which will contain individuals who are at clinical levels of mental health problems, and also importantly those who are at pre-clinical levels and those who may never develop a mental health problem. By taking a whole-population approach, these different trajectories can be unpicked; this will help to identify what factors are associated with vulnerability or resilience to childhood mental health problems and aid in the development of accurate early identification models. Moreover, utilising SAIL, we were also able to link three social care datasets to health and education data, rather than simply inferring social care involvement via other data sources, thus building on the work of Downs et al. (2019).

The use of the SAIL Databank is in itself a strength of the present study when trying to develop early identification tools for young people's mental health problems. Unlike a research database, SAIL contains rich routinely collected data with minimal data pre-processing or cleaning. This is a particular strength for this study because we needed the data to be as close to real-world routinely collected data as possible, with the aim for model performance to be preserved when translating it into clinical settings. One notable difference, however, between the SAIL data and that available to professionals in routine care is the census (i.e. cross-sectional) nature of the SAIL social care data rather than it being structured by episode, as would be the case in the raw social care data in routine practice. A second difference is the omission of unstructured, free-text in all of the SAIL datasets. In CADRE, we plan to include unstructured text in a pseudonymised form in the future. Pseudonymisation of free-text will be carried out using Clinical Records Anonymisation and Text Extraction (CRATE) software (Cardinal, 2017). This unstructured text could be interrogated using natural language processing (NLP) algorithms. There is some evidence to suggest that the inclusion of risk factors from free-text may enhance model performance (see for example the use of NLP in a crisis prediction tool known as the Management and Supervision Tool [MaST]; NHS, n.d.). However, it could also be that model performance is not as greatly improved by the inclusion of free-text information as one might anticipate; for example, Clayton et al.'s analysis in children's social care found that models including text data did not perform better than the models using structured data alone. Moreover, these models tended to learn patterns which did not generalise more than the models including structured data alone (Clayton et al., 2020). In our future steps, we will explore ways to enhance model performance such as the inclusion of more risk factors, including those captured from free-text data. With regards to future work in the CADRE database specifically, the advantage of our approach is that it uses real data from sites. Moreover, there is minimal data cleaning, equivalent to that which will happen in any local linked dataset. The data will be much the same as the data available via Secure Data Environments (SDEs) of the future. SDEs are currently being created across England and will be available for research. Local shared care records are also being created and will enable future tools to be run directly in the care environment. With regards to the present work in SAIL, the current models are simply prototypes – the first step in a long process to build models for early identification. The next stage would be to improve model performance, followed by external validation in other databases. At that stage, we would look to translate the "models" into a "digital tool" (which includes consideration of implementation infrastructure and governance). A variety of implementation options exist, such as a flag on GP records or social care records; the algorithm being run on local SDE datasets and results

reviewed by a multidisciplinary team including social care; the tool being run via a program or app that is linked to the local SDE dataset; and so on. However, how this will actually look in practice will be heavily informed by consultation with our data providing stakeholders and patient and public advisory panel.

Through this work, we were also able to develop vital collaborations with other teams, such as the Adolescent Mental Health Data Platform (ADP), strengthening and building capacity in our analytical work. Dr Friedmann provided a wealth of expertise and familiarity with the SAIL databases, linking together datasets which have not previously been linked together in SAIL. Without this working knowledge, data linkage alone could have taken the majority of the project time. This highlighted the resources required to undertake extensive administrative data linkage projects. Furthermore, the ADP, led by Professor John, collaborated with us, generously sharing validated code lists for mental health problems (Economou et al., 2012; John et al., 2021, 2018, 2016; McGregor et al., 2010; Rees et al., 2022; Wood et al., 2019). This built capacity for us to robustly measure mental health problems, making use of considerable prior research to validate these code lists so that they incorporated diagnostic codes and proxies for mental health problems (such as prescriptions).

Finally, this process has led to invaluable learning which can be applied to the build of the CADRE database. This learning was enabled by virtue of the fact that SAIL is characteristically similar to the database and Trusted Research Environment (TRE) we are building with CADRE. This learning will expedite the database-building process, as well as acting as a crucial external validation step for any early identification models we build in CADRE and similar TREs we are building in Birmingham and Essex. Through analysis in SAIL, we have been able to understand the importance of allowing significant time and consideration for data cleaning and processing. Through the risk factor mapping exercise, we have seen the importance of having clear and comprehensive meta-data, learnt about the different spectrum of measurability (from “directly measurable” to “derivable” to “measurable by proxy”), and seen the value of multi-agency data from health and local authorities. Through developing prototype risk prediction models, we have understood the significant importance of having timestamped data in order to truly model longitudinal change. Alongside analytical learning, we have gained important knowledge about information governance structures surrounding such a vast data resource as SAIL. This learning can be translated and modified as we develop CADRE, ensuring that this database is a valuable resource for clinicians and researchers, and leads to improvements in the lives of young people and the public.

6. Conclusions

It has long been understood that a wide variety of factors impact young people's mental health, with models such as Bronfenbrenner's ecological model and the bio–psycho–social model as just two examples of this recognition of the importance of factors beyond the individual. In order to truly understand complex issues such as mental health problems, it is important to consider the system or systems a young person is part of. Data linkage is a valuable approach to explore this issue. By bringing together administrative data from the organisations and services with which young people interact, mental health problems and their associated risk factors can more comprehensively be understood. In addition, this linkage of multi-agency data offers a promising way of developing early identification tools because early warning signs for mental health problems which may be missed in single-agency data can be combined together, leading to a stronger signal amid the noise, and avoiding young people falling through the gaps. Early identification of potential problems means that young people and their families, who may typically have to wait in distress for significant periods of time, can instead be identified earlier and offered timely and proportionate support. Furthermore, as more robust early identification tools are developed, staff in contexts such as social care can use them as a decision-making adjunct to help them identify young people who may have additional needs and to support smoother care pathways for young people and their families.

7. Recommendations and implications

Implications

This project has provided evidence to support the hypothesis that multi-agency data has value when attempting to measure a breadth of risk factors for childhood mental health problems and develop early identification tools. Furthermore, it has highlighted the added value of the inclusion of social care data in linked data projects focused on health.

Work in SAIL has also helped us to develop methodologies which can be translated to other projects and settings. These methodologies relate to:

1. How best to accurately measure mental health risk factors in administrative datasets.
2. How linked administrative health and social care data can be used to identify patterns indicating risk which may otherwise be missed when looking at single data sources.

This work to develop prototype early identification models in children's social care settings provides a solid foundation for future work, and our findings indicate it is worth continuing to refine the prototype models. Refinements include the addition of more risk factors, for example those extracted by natural language processing (NLP) methods, and also the inclusion of longitudinal data, additional datasets (such as the newly acquired Welsh Adoption Dataset) and larger datasets. Replication in different geographically bound populations will also be important for model refinement and evaluation of generalisability; the development of a network of Trusted Research Environments (TREs) across Cambridge, Birmingham and Essex will help to facilitate this.

Prediction models for mental health problems will allow mental health 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 of mental health problems and associated risk factors provides valuable information to commissioners to inform service planning and resource allocation based on the needs of the local population.

Finally, data on the levels of mental health 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 and healthcare will allow us to provide a more accurate estimation of this unrecognised mental health need in social care settings.

Recommendations and next steps

With respect to SAIL, we have successfully applied for funding and gained approval from SAIL to extend our data access. This will enable us to continue work in the linked database we have created as part of this project. Measurement of mental health problems will be repeated in the linked dataset, in order to see if different cases are identified through taking

a multi-agency approach to mental health condition measurement. Measurement of the unrecognised mental health need in children's social care will be carried out using the linked data. The prototype early identification model will also be refined, incorporating risk factors from the wealth of datasets. This will allow us to understand which datasets are critically important for predictive power and which variables are of particular importance for inclusion in the models. We will also visualise the prevalence of mental health problems and associated risk factors across Wales which will be a valuable resource for service commissioners.

With respect to CADRE, we will replicate the work from SAIL to understand if similar patterns are found in Cambridgeshire and Peterborough, UK. This will involve mapping equivalent risk factors to CADRE meta-data, measuring these risk factors and incorporating them into risk prediction models, helping us to understand if the models retain accuracy when translated into another database and population.

With respect to the wider Timely project, we recognise the necessity of including data from diverse populations and regions if we are to build accurate and fair early identification tools for childhood mental health problems. As such, with funding from the Alan Turing Institute and HDR UK, we have begun the process of developing data-sharing agreements and governance structures for a network of Trusted Research Environments (TREs) with federated analytics. In this network, there will be data from the populations of Cambridgeshire, Peterborough, Birmingham and Essex. Weaved throughout this workstream has been active engagement with advisory panels of young people, parents and carers, to ensure that the Timely project makes use of people's data in a way which is acceptable, desirable and understandable.

References

Allen, G. (2011) *Early intervention: The next steps - An independent report to Her Majesty's Government*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/284086/early-intervention-next-steps2.pdf [Accessed 2nd February 2023].

Berridge, D., Luke, N., Sebba, J., Strand, S., Cartwright, M., Staples, E., … O'Higgins, A. (2020) Children in need and children in care: Educational attainment and progress. http://www.bristol.ac.uk/policybristol/policy-briefings/children-in-need-and-in-care-education-progress/ [Accessed 2nd February 2023].

Bronfenbrenner, U. (1995) Developmental ecology through space and time: A future perspective. In P. Moen, G. Elder Jr. & K. Lüscher (Eds.), *Examining lives in context: Perspectives on the ecology of human development* (pp. 619–647). Washington DC: American Psychological Association.

Bronfenbrenner, U. & Morris, P. A. (2006) The bioecological model of human development. In R. M. Lerner & W. E. Damon (Eds.), *Handbook of child psychology: Vol 1, theoretical models of human development* (pp. 793–828). West Sussex: Wiley.

Cardinal, R. N. (2017) Clinical records anonymisation and text extraction (CRATE): An open-source software system. *BMC Medical Informatics and Decision Making*. 17 (1), 1–12. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0437-1.

Care Leavers' Association. (2017) *Caring for better health: An investigation into the health needs of care leavers*. https://www.careleavers.com/wp-content/uploads/2017/12/Caring-for-Better-Health-Final-Report.pdf [Accessed 2nd February 2023].

Clayton, V., Sanders, M., Schoenwald, E., Surkis, L. & Gibbons, D. (2020) *Machine learning in children's services: Technical report*. What Works For Children's Social Care. https://whatworks-csc.org.uk/wp-content/uploads/WWCSC_technical-_report_machine_learning_in_childrens_services_does_it_work_Sep_2020.pdf [Accessed 2nd February 2023].

Davis, K. A. S., Sudlow, C. L. M. & Hotopf, M. (2016) Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses. *BMC Psychiatry*. 16 (1), 263. https://doi.org/10.1186/s12888-016-0963-x.

Department for Education [DfE]. (2020) *Characteristics of children in need 2020*. https://explore-education-statistics.service.gov.uk/find-statistics/characteristics-of-children-in-need [Accessed 2nd February 2023].

Department for Education [DfE]. (2021) *Children looked after in England including adoptions*. Retrieved from https://explore-education-statistics.service.gov.uk/find-statistics/children-looked-after-in-england-including-adoptions/2020. [Accessed 2nd February 2023].

Department of Health and Social Care & Department for Education [DHSC & DfE]. (2018) *Transforming children and young people's mental health provision: A green paper and next steps*. https://www.gov.uk/government/consultations/transforming-children-and-young-peoples-mental-health-provision-a-green-paper [Accessed 2nd February 2023].

Downs, J., Gilbert, R., Hayes, R. D., Hotopf, M. & Ford, T. (2017) Linking health and education data to plan and evaluate services for children. *Archives of Disease in Childhood*. 102 (7), 599–602. https://doi.org/10.1136/archdischild-2016-311656.

Downs, J. M., Ford, T., Stewart, R., Epstein, S., Shetty, H., Little, R., … Hayes, R. (2019) 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*. 9 (1). https://doi.org/10.1136/bmjopen-2018-024355.

Economou, A., Grey, M., McGregor, J., Craddock, N., Lyons, R. A., Owen, M. J., … Lloyd, K. (2012) The health informatics cohort enhancement project (HICE): Using routinely collected primary care data to identify people with a lifetime diagnosis of psychotic disorder. *BMC Research Notes*. 5, 95. https://doi.org/10.1186/1756-0500-5-95.

Fong, H., French, B., Rubin, D. & Wood, J. N. (2015). Mental health services for children and caregivers remaining at home after suspected maltreatment. *Children and Youth Services Review*. 58, 50–59. https://doi.org/10.1016/j.childyouth.2015.08.010.

Ford D. V., Jones K. H., Verplancke, J. P., Lyons, R. A., John, G., Brown, G., Brooks, C. J., Thompson, S., Bodger, O., Couch, T. & Leake K. (2009) The SAIL Databank: Building a national architecture for e-health research and evaluation. *BMC Health Services Research*. 9 (1), 1–12. https://doi.org/10.1186/1472-6963-9-157.

Grath-Lone, L. M., Libuy, N., Blackburn, R., Harron, K., Etoori, D. & Gilbert, R. (2021) 921 The education and child health insights from linked data (ECHILD) database: A newly linked, de-identified health-education data resource. *Archives of Disease in Childhood*. 106 (Suppl 1), A163 LP-A164. https://doi.org/10.1136/archdischild-2021-rcpch.284.

John, A., Friedmann, Y., DelPozo-Banos, M., Frizzati, A., Ford, T. & Thapar, A. (2021) Association of school absence and exclusion with recorded neurodevelopmental disorders, mental disorders, or self-harm: A nationwide, retrospective, electronic cohort study of children and young people in Wales, UK. *The Lancet Psychiatry*. 9 (1), 23–34. https://doi.org/10.1016/S2215-0366(21)00367-9.

John, A., McGregor, J., Jones, I., Lee, S. C., Walters, J. T. R., Owen, M. J., … Lloyd, K. (2018) Premature mortality among people with severe mental illness: New evidence from linked primary care data. *Schizophrenia Research*. 199, 154–162. https://doi.org/10.1016/j.schres.2018.04.009.

John, A., McGregor, J., Fone, D., Dunstan, F., Cornish, R., Lyons, R. A. & Lloyd, K. R. (2016) Case-finding for common mental disorders of anxiety and depression in primary care: An external validation of routinely collected data. *BMC Medical Informatics and Decision Making*. 16 (1), 1–10. https://doi.org/10.1186/s12911-016-0274-7.

Jones, K. H., Ford, D. V., Thompson, S. & Lyons, R. A. (2019) A profile of the SAIL Databank on the UK secure research platform. *International Journal of Population Data Science*. 4 (2). https://doi.org/10.23889/ijpds.v4i2.1134.

Lee, A., Elliott, M., Scourfield, J., Bedston, S., Broadhust, K., Ford, D. V & Griffiths, L. J. (2022) Data resource: Children receiving care and support and children in need, administrative records in Wales. *International Journal of Population Data Science*. 7 (1). https://doi.org/10.23889/ijpds.v7i1.1694.

Lyons, R. A., Jones, K. H., John, G., Brooks, C. J., Verplancke, J. P., Ford, D. V., … Leake, K. (2009) The SAIL Databank: Linking multiple health and social care datasets. *BMC Medical Informatics and Decision Making*. 9 (1), 1–8. https://doi.org/10.1186/1472-6947-9-3.

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

Malone, B., Garcia-Duran, A. & Niepert, M. (2018) Learning representations of missing data for predicting patient outcomes. https://doi.org/10.48550/arXiv.1811.04752.

Matuszak, J. & Piasecki, M. (2012) Inter-rater reliability in psychiatric diagnosis. *Psychiatric Times*. 29 (10). https://www.psychiatrictimes.com/view/inter-rater-reliability-psychiatric-diagnosis. [Accessed 2nd February 2023].

McGregor, J., Brooks, C., Chalasani, P., Chukwuma, J., Hutchings, H., Lyons, R. A. & Lloyd, K. (2010) The Health Informatics Trial Enhancement Project (HITE): Using routinely collected primary care data to identify potential participants for a depression trial. *Trials*. 11 (1), 39. https://doi.org/10.1186/1745-6215-11-39.

Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D. & Danese, A. (2022) Clinical prediction models in psychiatry: A systematic review of two decades of progress and challenges. *Molecular Psychiatry*. 27 (6), 2700–2708. https://doi.org/10.1038/s41380-022-01528-4.

NHS. (n.d.) A caseload management and supervision tool for community mental health services. https://transform.england.nhs.uk/key-tools-and-info/digital-playbooks/mental-health-digital-playbook/a-caseload-management-and-supervision-tool-for-community-mental-health-services/. [Accessed 2nd February 2023].

Nicholls, S. G., Langan, S. M. & Benchimol, E. I. (2017) Routinely collected data: The importance of high-quality diagnostic coding to research. *Canadian Medical Association Journal*. 189 (33), E1054–E1055. https://doi.org/10.1503/cmaj.170807.

Rees, S., Watkins, A., Keauffling, J. & John, A. (2022) Incidence, mortality and survival in young people with co-occurring mental disorders and substance use: A retrospective linked routine data study in Wales. *Clinical Epidemiology*. 14, 21–38. https://doi.org/10.2147/CLEP.S325235.

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A. & Kupfer, D. J. (2013) DSM-5 field trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry*. 170 (1), 59–70. https://doi.org/10.1176/appi.ajp.2012.12070999.

Rocheteau, E., Tong, C., Veličković, P., Lane, N. & Liò, P. (2021) Predicting patient outcomes with graph representation learning. https://doi.org/10.48550/arXiv.2101.03940.

Sanders, R. (2020) *Care experienced children and young people's mental health*. Glasgow, Iriss. https://www.iriss.org.uk/resources/outlines/care-experienced-children-and-young-peoples-mental-health [Accessed 2nd February 2023].

Sze, V., Chen, Y.-H., Yang, T.-J. & Emer, J. S. (2017) Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*. 105 (12), 2295–2329. https://doi.org/10.1109/jproc.2017.2761740.

Thayer, D., Rees, A., Kennedy, J., Collins, H., Harris, D., Halcox, J., ... & Brooks, C. (2020) Measuring follow-up time in routinely-collected health datasets: Challenges and solutions. *PLoS One*. 15 (2), e0228545. https://doi.org/10.1371/journal.pone.0228545.

The Child Safeguarding Practice Review Panel. (2021) *Annual report 2020: Patterns in practice, key messages and 2021 work programme*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/984767/The_Child_Safeguarding_Annual_Report_2020.pdf. [Accessed 2nd February 2023].

Tonekaboni, S., Joshi, S., McCradden, M. D. & Goldenberg, A. (2019) What clinicians want: Contextualizing explainable machine learning for clinical end use. Proceedings of the 4th Machine Learning for Healthcare Conference. *PMLR*. 106, 359–380. https://proceedings.mlr.press/v106/tonekaboni19a.html. [Accessed 2nd February 2023].

What Works for Children's Social Care. (2016) *Study review: Mental health care interventions for children looked after*. https://whatworks-csc.org.uk/evidence/evidence-store/intervention/mental-health-care-interventions-for-children-looked-after/. [Accessed 2nd February 2023].

Wood, S., Marchant, A., Allsopp, M., Wilkinson, K., Bethel, J., Jones, H. & John, A. (2019) Epidemiology of eating disorders in primary care in children and young people: A Clinical Practice Research Datalink study in England. *BMJ Open*. 9 (8), e026691. http://dx.doi.org/10.1136/bmjopen-2018-026691.

Xiao, C., Choi, E. & Sun, J. (2018) Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*. 25 (10), 1419–1428. https://doi.org/10.1093/jamia/ocy068.

Appendices

Appendix A

Table A1: Qualitative table to show which risk factors we consider to be measurable in each database

Database name	Directly measurable risk factors	Total directly measurable	Derivable risk factors	Total derivable	Total measurable
Dataset type: Demographics					
Welsh Demographic Service Dataset (WDSD)	Month of birth; Sex (biological)	2	Area deprivation (area code)	1	**3**
Annual District Birth Extract (ADBE)	Birth weight; Urbanicity	2	Primary caregiver(s) unemployment	1	**3**
Annual District Death Extract (ADDE)	N/A	0	Acute stress disorder (as a predictor of further mental health problems); Acute stress symptoms (anxiety, avoidance or depression – as a	32	**32**

			predictor of further mental health problems); Allergies (e.g. non-lgE-mediated food allergies, pollen allergies); Anaemia; Anxiety (as a predictor of further mental health problems); Asthma; Autoimmune disorders (e.g. rheumatoid arthritis); Chronic (long lasting) infection (e.g. Lyme disease, periodontal disease); Chronic (long lasting) gastric ill-health (e.g. inflammatory bowel disease); Chronic (long lasting) reflux or indigestion; Chronic inflammation; Congenital malformations; Diabetes; Eczema; Hypoxia (at birth); Increased panic attacks; Inflammatory diseases (e.g. Lyme disease); Intellectual disability; Irritable Bowel Syndrome (IBS); Long-term antibiotic use; Low levels of B vitamins; Low levels of folate; Low levels of vitamin D; Low serum ferritin; Low serum vitamin D; Obesity/overweight; Prolonged duration of a physical health condition; Repeated infections; Severe health condition; Sleep disorder; Thyroid disease; Traumatic brain injury		

Dataset type: Education					
Pre-16 Education Attainment (EDUW)	Ethnicity; Out-of-school discipline (e.g. suspension and expulsion); Participation in the Free-Lunch Programme; School exclusions; Special Educational Needs (SEN)	5	Poor educational attainment; Poor school attendance; School-level deprivation (e.g. proportion eligible for free school meals)	3	**8**
Dataset type: Social care					
Child In Need – Wales (CINW)	Being a looked after child (LAC); Child in need (CIN) status; Child protection record; Domestic violence; Ethnicity; Hearing impairment causing disability (e.g. deafness); Household alcohol abuse; Household drug abuse; Household mental illness; Involvement in criminal justice system; Neuro-developmental conditions (e.g. autism); Non-prescription drug use; Out-of-school discipline (e.g. suspension and expulsion); Physical disability; Primary caregiver(s) chronic (long lasting) illness; Primary caregiver(s) mental health problems; Problems with memory; Psychiatric history; School exclusions; Severe illness in family; Visual impairment	21	Emotional neglect; Failure to attend three or more planned health or social care appointments; Physical neglect	3	**24**

	causing disability (e.g. blindness/partial sightedness)				
Children Receiving Care and Support (CRCS)	Being a looked after child (LAC); Child in need (CIN) status; Child protection record; Domestic violence; Emotional neglect; Emotional, psychological or verbal abuse; Experiencing financial abuse; Hearing impairment causing disability (e.g. deafness); Household alcohol abuse; Household drug abuse; Household mental illness; Involvement in criminal justice system; Neuro-developmental conditions (e.g. autism); Non-prescription drug use; Physical abuse; Physical disability; Physical neglect; Problems with memory; Primary caregiver(s) chronic (long lasting) illness; Primary caregiver(s) mental health problems; Psychiatric history; Severe illness in family; Sexual abuse; Visual impairment causing disability (e.g. blindness/partial sightedness)	24	Failure to attend three or more planned health or social care appointments	1	**25**

Looked After Children – Wales (LACW)	Being an unaccompanied asylum seeker; Being a looked after child (LAC)	2	Emotional neglect; Physical neglect	2	**4**
Dataset type: General practice/primary care					
GP Primary Care (WLGP)	N/A	0	Acute stress disorder (as a predictor of further mental health problems); Acute stress symptoms (anxiety, avoidance or depression – as a predictor of further mental health problems); Allergies (e.g. non-lgE-mediated food allergies, pollen allergies); Anaemia; Anxiety (as a predictor of further mental health problems); Asthma; Autoimmune disorders (e.g. rheumatoid arthritis); Birth length; Chronic (long lasting) infection (e.g. Lyme disease, periodontal disease); Chronic (long lasting) gastric ill-health (e.g. inflammatory bowel disease); Chronic (long lasting) reflux or indigestion; Chronic inflammation; Congenital malformations; Diabetes; Disease history; Eczema; Hypoxia (at birth); Increased panic attacks; Inflammatory diseases (e.g. Lyme disease); Intellectual disability;	33	**33**

			Irritable Bowel Syndrome (IBS); Long-term antibiotic use; Low levels of B vitamins; Low levels of folate; Low levels of vitamin D; Low serum ferritin; Low serum vitamin D; Obesity/overweight; Prolonged duration of a physical health condition; Repeated infections; Severe health condition; Sleep disorder; Thyroid disease; Traumatic brain injury		
Dataset type: Specialist or acute healthcare					
Critical Care Dataset (CCDS)	Urbanicity	1	Disease history; Repeat hospitalisation	2	**3**
Emergency Department Dataset (EDDS)	N/A	0	Acute stress disorder (as a predictor of further mental health problems); Acute stress symptoms (anxiety, avoidance or depression – as a predictor of further mental health problems); Allergies (e.g. non-lgE-mediated food allergies, pollen allergies); Anaemia; Anxiety (as a predictor of further mental health problems); Asthma; Autoimmune disorders (e.g. rheumatoid arthritis);	33	**33**

			Chronic (long lasting) infection (e.g. Lyme disease, periodontal disease); Chronic (long lasting) gastric ill-health (e.g. inflammatory bowel disease); Chronic (long lasting) reflux or indigestion; Chronic inflammation; Congenital malformations; Diabetes; Disease history; Eczema; Hypoxia (at birth); Increased panic attacks; Inflammatory diseases (e.g. Lyme disease); Intellectual disability; Irritable Bowel Syndrome (IBS); Long-term antibiotic use; Low levels of B vitamins; Low levels of folate; Low levels of vitamin D; Low serum ferritin; Low serum vitamin D; Obesity/overweight; Prolonged duration of a physical health condition; Repeated infections; Severe health condition; Sleep disorder; Thyroid disease; Traumatic brain injury		
National Community Child Health (NCCH)	5-min Apgar score <7; Birth weight; Maternal smoking during pregnancy; Not breastfed; Premature birth	5	Birth length; Body mass Index (BMI); Disease history; Failure to attend three or more planned health or social care appointments; Global developmental delay; Obesity/overweight	6	**11**

NHS 111 Call Data (NHSO)	Ethnicity	1	Investigations by multiple services suggestive of suffering from medically unexplained symptoms; Three or more presentations to emergency services within a year	2	**3**
NHS Hospital Outpatients (OPDW)	N/A	0	Acute stress disorder (as a predictor of further mental health problems); Acute stress symptoms (anxiety, avoidance or depression – as a predictor of further mental health problems); Allergies (e.g. non-IgE-mediated food allergies, pollen allergies); Anaemia; Anxiety (as a predictor of further mental health problems); Asthma; Autoimmune disorders (e.g. rheumatoid arthritis); Chronic (long lasting) infection (e.g. Lyme disease, periodontal disease); Chronic (long lasting) gastric ill-health (e.g. inflammatory bowel disease); Chronic (long lasting) reflux or indigestion; Chronic inflammation; Congenital malformations; Diabetes; Disease history; Eczema; Hypoxia (at birth); Increased panic attacks; Inflammatory diseases (e.g. Lyme disease); Intellectual disability; Irritable Bowel Syndrome (IBS);	33	**33**

			Long-term antibiotic use; Low levels of B vitamins; Low levels of folate; Low levels of vitamin D; Low serum ferritin; Low serum vitamin D; Obesity/overweight; Prolonged duration of a physical health condition; Repeated infections; Severe health condition; Sleep disorder; Thyroid disease; Traumatic brain injury		
Outpatient Referral Dataset (OPRD)	Urbanicity	1	Disease history	1	**2**
Patient Episode Database – Wales (PEDW)	N/A	0	Acute stress disorder (as a predictor of further mental health problems); Acute stress symptoms (anxiety, avoidance or depression – as a predictor of further mental health problems); Allergies (e.g. non-lgE-mediated food allergies, pollen allergies); Anaemia; Anxiety (as a predictor of further mental health problems); Asthma; Autoimmune disorders (e.g. rheumatoid arthritis); Chronic (long lasting) infection (e.g. Lyme disease, periodontal disease);	33	**33**

			Chronic (long lasting) gastric ill-health (e.g. inflammatory bowel disease); Chronic (long lasting) reflux or indigestion; Chronic inflammation; Congenital malformations; Diabetes; Disease history; Eczema; Hypoxia (at birth); Increased panic attacks; Inflammatory diseases (e.g. Lyme disease); Intellectual disability; Irritable Bowel Syndrome (IBS); Long-term antibiotic use; Low levels of B vitamins; Low levels of folate; Low levels of vitamin D; Low serum ferritin; Low serum vitamin D; Obesity/overweight; Prolonged duration of a physical health condition; Repeated infections; Severe health condition; Sleep disorder; Thyroid disease; Traumatic brain injury		
Substance Misuse Dataset (SMDS)	Being a young parent; Heavy alcohol use; Homelessness; Homelessness (young person has left home); Non-prescription drug use; Serving in the military; Unemployment of the individual	7	Disease history; Nitrous oxide use; Poor school attendance; Prescription drug abuse; Psychiatric history; Severe health condition	6	**7**

Wales Results Reporting Service (WRRS)	N/A	0	Acute stress disorder (as a predictor of further mental health problems); Acute stress symptoms (anxiety, avoidance or depression – as a predictor of further mental health problems); Allergies (e.g. non-lgE-mediated food allergies, pollen allergies); Anaemia; Anxiety (as a predictor of further mental health problems); Asthma; Autoimmune disorders (e.g. rheumatoid arthritis); Birth length; Chronic (long lasting) infection (e.g. Lyme disease, periodontal disease); Chronic (long lasting) gastric ill-health (e.g. inflammatory bowel disease); Chronic (long lasting) reflux or indigestion; Chronic inflammation; Congenital malformations; Diabetes; Disease history; Eczema; Hypoxia (at birth); Increased panic attacks; Inflammatory diseases (e.g. Lyme disease); Intellectual disability; Investigations by multiple services suggestive of suffering from medically unexplained symptoms; Irritable Bowel Syndrome (IBS); Long-term antibiotic use; Low levels of B vitamins; Low levels of folate; Low levels of vitamin D; Low serum	36	**36**

			ferritin; Low serum vitamin D; Obesity/overweight; Prolonged duration of a physical health condition; Repeated infections; Severe health condition; Sleep disorder; Three or more presentations to emergency services within a year; Thyroid disease; Traumatic brain injury		
Maternity Indicators Dataset (MIDS)	5-min Apgar score <7; Age of parents at time of giving birth; Birth weight; Family history of psychiatric disorders; Maternal obesity/overweight during pregnancy; Maternal smoking during pregnancy; Month of birth; Not breastfed; Premature birth; Primary caregiver(s) mental health problems; Smoking status of primary caregiver(s)	11	Family history of severe mental illness (e.g. psychosis); Maternal depression during pregnancy; Household mental illness	3	**14**

Appendix B

Table B1: Qualitative table to show which risk factors relating to under-served populations we consider to be measurable in each database

Database name	Directly measurable risk factors	Total directly measurable	Derivable risk factors	Total derivable	Total measurable
Dataset type: Demographics					
Welsh Demographic Service Dataset (WDSD)	N/A	0	Area deprivation (area code)	1	**1**
Annual District Birth Extract (ADBE)	Urbanicity	1	Primary caregiver(s) unemployment	1	**2**
Annual District Death Extract (ADDE)	N/A	0	Intellectual disability	1	**1**

Dataset type: Education					
Pre-16 Education Attainment (EDUW)	Ethnicity; Out-of-school discipline (e.g. suspension and expulsion); Participation in the Free-Lunch Programme; School exclusions; Special Educational Needs (SEN)	5	Poor school attendance; Poor educational attainment; School-level deprivation (e.g. proportion eligible for free school meals)	3	**8**
Dataset type: Social care					
Child In Need – Wales (CINW)	Being a looked after child (LAC); Child in need (CIN) status; Child protection record; Ethnicity; Hearing impairment causing disability (e.g. deafness); Household alcohol abuse; Household drug abuse; Household mental illness; Involvement in criminal justice system; Neuro-developmental conditions (e.g. autism); Non-prescription drug use; Out-of-school discipline (e.g. suspension and expulsion); Physical disability; Primary caregiver(s) mental health problems; School exclusions; Visual impairment causing disability (e.g. blindness/partial sightedness)	16	Failure to attend three or more planned health or social care appointments	1	**17**
Children Receiving Care and	Being a looked after child (LAC); Child in need (CIN) status; Child protection record; Hearing impairment causing disability (e.g. deafness); Household alcohol	13	Failure to attend three or more planned health or	1	**14**

Support (CRCS)	abuse; Household drug abuse; Household mental illness; Involvement in criminal justice system; Neuro-developmental conditions (e.g. autism); Non-prescription drug use; Physical disability; Primary caregiver(s) mental health problems; Visual impairment causing disability (e.g. blindness/partial sightedness)		social care appointments		
Looked After Children – Wales (LACW)	Being an unaccompanied asylum seeker; Being a looked after child (LAC)	2	N/A	0	**2**
Dataset type: General practice/primary care					
GP Primary Care (WLGP)	N/A	0	Intellectual disability	1	**1**
Dataset type: Specialist or acute healthcare					
Critical Care Dataset (CCDS)	Urbanicity	1	N/A	0	**1**
Emergency Department	N/A	0	Intellectual disability	1	**1**

Dataset (EDDS)					
National Community Child Health (NCCH)	N/A	0	Failure to attend three or more planned health or social care appointments	1	**1**
NHS 111 Call Data (NHSO)	Ethnicity	1	N/A	0	**1**
NHS Hospital Outpatients (OPDW)	N/A	0	Intellectual disability	1	**1**
Outpatient Referral Dataset (OPRD)	Urbanicity	1	N/A	0	**1**
Patient Episode Database – Wales (PEDW)	N/A	0	Intellectual disability	1	**1**

Substance Misuse Dataset (SMDS)	Being a young parent; Heavy alcohol use; Homelessness; Homelessness (young person has left home); Non-prescription drug use; Served in the military; Unemployment of the individual	7	Nitrous oxide use; Poor school attendance; Prescription drug abuse	3	**10**
Wales Results Reporting Service (WRRS)	N/A	0	Intellectual disability	1	**1**
Maternity Indicators Dataset (MIDS)	Primary caregiver(s) mental health problems	1	Household mental illness; Maternal depression during pregnancy	2	**3**

Appendix C

Table C1: Mental health diagnoses measured by dataset

Dataset	Diagnoses available (Y/N)	Coding system	Dataset size linked with our cohort	Distinct patients with at least 1 valid diagnosis (§ OR Cause of Death for ADDE)	Completeness of diagnoses (i.e. physical or mental health) (in %)	Prevalence of any mental health problem (%)	Depression (%)	Anxiety (%)	^Mood disorder (inc. depression, anxiety) (%)
Patient Episode Database for Wales (PEDW)	Y	ICD-10	1,113,776	635,481	57.06	30,350 (4.78)	10,451 (1.64)	12,615 (1.99)	17,340 (2.73)
NHS Hospital Outpatients (OPDW)	Y	ICD-10	658,890	36,195	5.49	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
GP Primary Care – Audit (WLGP)	Y	READ codes	950,532	949,548	99.90	141,052 (14.85)	87,628 (9.23)	74,185 (7.81)	123,077 (12.96)
Outpatient Referrals from Primary Care (OPRD)	N	–	–	–	–	–	–	–	–
National Community Child Health Database (NCCH)	N	–	–	–	–	–	–	–	–
Emergency Department Dataset (EDDS)	Y	ICD-10	710,752	693,410	97.56	18 (~0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Critical Care Dataset (CCDS)	N	–	–	–	–	–	–	–	–

Wales Results Reporting Service (WRRS)	Y	READ codes	876,039	524,408	59.86	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Substance Misuse Dataset (SMDS)	N	–	–	–	–	–	–	–	–
Maternity Indicators Dataset (MIDS)	N	–	–	–	–	–	–	–	–
NHS 111 Call Data (NHSO)	N	–	–	–	–	–	–	–	–
§Annual District Death Extract (ADDE)	Y	ICD-10	2,400	2,400	100	32 (1.33)	0 (0.00)	0 (0.00)	0 (0.00)

Table C1 (continued): Mental health diagnoses measured by dataset

Datasets	Alcohol misuse (%)	Drug misuse (%)	^Substance misuse (inc. alcohol, drug) (%)	Bipolar (%)	Schizophrenia (%)	^Severe mental illness (inc. bipolar, schizophrenia) (%)	Conduct disorder (%)	Eating disorder (%)	Self-harm (intentional; 10+ years) (%)	Self-harm (undetermined Intent; 15+ years) (%)
Patient Episode Database for Wales (PEDW)	5,468 (0.86)	6,820 (1.07)	10,461 (1.65)	462 (0.07)	424 (0.07)	836 (0.13)	829 (0.13)	1,386 (0.22)	11,440 (1.80)	667 (0.10)
NHS Hospital Outpatients (OPDW)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
GP Primary Care – Audit (WLGP)	3,844 (0.40)	5,608 (0.59)	8,636 (0.91)	360 (0.04)	759 (0.08)	1,086 (0.11)	7,258 (0.76)	10,643 (1.12)	19,527 (2.06)	144 (0.02)

Outpatient Referrals from Primary Care (OPRD)	–	–	–	–	–	–	–	–	–	–
National Community Child Health Database (NCCH)	–	–	–	–	–	–	–	–	–	–
Emergency Department Dataset (EDDS)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	18 (~0.00)	0 (0.00)	0 (0.00)
Critical Care Dataset (CCDS)	–	–	–	–	–	–	–	–	–	–
Wales Results Reporting Service (WRRS)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Substance Misuse Dataset (SMDS)	–	–	–	–	–	–	–	–	–	–
Maternity Indicators Dataset (MIDS)	–	–	–	–	–	–	–	–	–	–
NHS 111 Call Data (NHSO)	–	–	–	–	–	–	–	–	–	–
§Annual District Death Extract (ADDE)	✧	✧	32 (1.33)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)

Notes: Prevalence measurements were completed using all available data from each dataset for our cohort so will not be comparable between datasets for the population because some datasets have been in existence for longer than others, so naturally contain more diagnostic codes. In a next step, we will explore a specific time period in order to understand the prevalence by population at a given period in time, rather than prevalence by dataset as currently presented.

Prevalence measurements will also be repeated in the linked dataset when enhanced computing power is provided by SAIL; this will avoid duplicate counting of diagnoses for the same individual in multiple datasets (i.e. where a diagnosis for an individual is recorded in more than one dataset rather than being true cases of multiple diagnoses for multiple individuals).

§ Diagnostic codes measured in the Annual District Death Extract (ADDE) relate specifically to recorded cause of death. Of note, suicide was not one of the conditions we measured so the ADDE row of data should be read with this in mind as it may be an underrepresentation of mental health diagnostic codes.

^ “Mood disorders” codes consist of depression and/or anxiety diagnostic codes. Since some individuals may have depression and/or anxiety codes in their records, the sum of these two conditions individually does not sum to the number in the aggregated “Mood disorders” column; in the “Mood disorders” where conditions are counted once if an individual has depression and/or anxiety. The same logic applies for “Substance misuse” that comprises alcohol misuse and drug misuse, and for “Severe mental illness” that comprises bipolar, schizophrenia and other psychotic disorders.

✧ The exact numbers of alcohol and drug misuse are not individually reported due to confidentiality issues (i.e. one of them is <5 individuals). Aggregated results are reported in the “Substance misuse” column.

Appendix D

Table D1: Odds ratios for interpretable logistic regression model*

Risk factor	Odds ratio
Apgar 1-Minute Score: Continuous Value	0.95
Apgar 1-Minute Score: Unknown	1.198
Apgar 5-Minute Score: Continuous Value	1.001
Apgar 5-Minute Score: Unknown	1.168
Asylum Seeker Status: Asylum Seeker	0.761
Asylum Seeker Status: Not Seeker	1.277
Autistic Spectrum Disorder Status: Autistic	2.258
Autistic Spectrum Disorder Status: Not Autistic	0.83
Autistic Spectrum Disorder Status: Unknown	0.519
Birth Weight: Continuous Value	1.036
Birth Weight: Unknown	0.844
Breastfeed Status (8 weeks): Breastfed	0.99
Breastfeed Status (8 weeks): Not Breastfed	0.953
Breastfeed Status (8 weeks): Unknown	1.03
Breastfeed Status (Birth): Breastfed	1.059
Breastfeed Status (Birth): Not Breastfed	1.067
Breastfeed Status (Birth): Unknown	0.859
Category of Need: Socially Unacceptable Behaviour	1.211
Category of Need: Absent Parenting	0.723
Category of Need: Abuse or Neglect	0.786
Category of Need: Adoption Disruption	1.825
Category of Need: Child’s Disability or Illness	0.994
Category of Need: Family Dysfunction	0.932
Category of Need: Family in Acute Stress	1.281

Category of Need: Low Income	0.471
Category of Need: Parental Disability or Illness	1.384
Child Protection Register Status: Not Registered	1.019
Child Protection Register Status: Registered	0.953
Dental Check Status: Not Receiving Checks	0.734
Dental Check Status: Receiving Checks	1.254
Dental Check Status: Unknown	1.055
Diagnostic Code: Intentional self-poisoning by drugs, medicaments and biological substances	1.86
Diagnostic Code: Other acute lower respiratory infections	0.956
Diagnostic Code: Other disorders of ear, Intraoperative and postprocedural complications and disorders of ear and mastoid process, not elsewhere classified	0.761
Diagnostic Code: Accidental poisoning by and exposure to noxious substances	0.979
Diagnostic Code: Acute upper respiratory infections	1.08
Diagnostic Code: Burns and corrosions of multiple and unspecified body regions, Frostbite, Poisoning by, adverse effect of and underdosing of drugs, medicaments and biological substances	1.269
Diagnostic Code: Chronic lower respiratory diseases	1.234
Diagnostic Code: Complications of labor and delivery	1.545
Diagnostic Code: Complications of surgical and medical care, not elsewhere classified	0.845
Diagnostic Code: Congenital malformations of the circulatory system	0.722
Diagnostic Code: Congenital malformations of the urinary system	0.844
Diagnostic Code: Diseases of oesophagus, stomach and duodenum	0.877
Diagnostic Code: Diseases of external ear, Diseases of middle ear and mastoid	0.832
Diagnostic Code: Diseases of male genital organs	1.243
Diagnostic Code: Diseases of oral cavity and salivary glands	1.175
Diagnostic Code: Encounters for other specific health care	1.145
Diagnostic Code: Episodic and paroxysmal disorders	1.109
Diagnostic Code: Exposure to inanimate mechanical forces	0.974
Diagnostic Code: General symptoms and signs	1.182
Diagnostic Code: General symptoms and signs	1.041

Diagnostic Code: Genetic carrier and genetic susceptibility to disease, Resistance to antimicrobial drugs, Oestrogen receptor status, Retained foreign body fragments, Hormone sensitivity malignancy status	1.345
Diagnostic Code: Haemorrhagic and hematological disorders of newborn	0.869
Diagnostic Code: Infections of the skin and subcutaneous tissue	1.017
Diagnostic Code: Infections specific to the perinatal period	1.033
Diagnostic Code: Influenza and pneumonia	0.931
Diagnostic Code: Injuries to the elbow and forearm	1.37
Diagnostic Code: Injuries to the head	0.972
Diagnostic Code: Injuries to the knee and lower leg	1.057
Diagnostic Code: Injuries to the wrist, hand and fingers	1.265
Diagnostic Code: Intestinal infectious diseases	0.952
Diagnostic Code: Maternal care related to the fetus and amniotic cavity and possible delivery problems	1.068
Diagnostic Code: Metabolic disorders, postprocedural endocrine and metabolic complications and disorders, not elsewhere classified	1.184
Diagnostic Code: Newborn affected by maternal factors and by complications of pregnancy, labor, and delivery, Disorders of newborn related to length of gestation and fetal growth, Abnormal findings on neonatal screening	1.054
Diagnostic Code: Noninfective enteritis and colitis, Other diseases of intestines	1.068
Diagnostic Code: Other and unspecified dermatitis	0.884
Diagnostic Code: Other diseases of the urinary system	0.976
Diagnostic Code: Other diseases of upper respiratory tract	1.01
Diagnostic Code: Other disorders originating in the perinatal period	0.862
Diagnostic Code: Other joint disorders, Dentofacial anomalies [including malocclusion] and other disorders of jaw	1.286
Diagnostic Code: Other maternal disorders predominantly related to pregnancy	1.038
Diagnostic Code: Other viral diseases, Mycoses	1.001
Diagnostic Code: Overexertion and strenuous or repetitive movements, Accidental exposure to other specified factors	0.93
Diagnostic Code: Persons encountering health services for examinations	0.991
Diagnostic Code: Persons encountering health services in circumstances related to reproduction	1.029
Diagnostic Code: Persons encountering health services in other circumstances	1.06

Diagnostic Code: Persons with potential health hazards related to family and personal history and certain conditions influencing health status	1.365
Diagnostic Code: Persons with potential health hazards related to family and personal history and certain conditions influencing health status	0.992
Diagnostic Code: Persons with potential health hazards related to socioeconomic and psychosocial circumstances	0.951
Diagnostic Code: Persons with potential health hazards related to socioeconomic and psychosocial circumstances, Do not resuscitate status, Blood type, Body mass index (BMI)	1.101
Diagnostic Code: Poisoning by, adverse effect of and underdosing of drugs, medicaments and biological substances	1.096
Diagnostic Code: Provisional assignment of new diseases of uncertain etiology or emergency use	1.232
Diagnostic Code: Respiratory and cardiovascular disorders specific to the perinatal period	1.046
Diagnostic Code: Sequelae of infectious and parasitic diseases, Bacterial and viral infectious agents, Other infectious diseases	1.012
Diagnostic Code: Slipping, tripping, stumbling and falls	1.069
Diagnostic Code: Slipping, tripping, stumbling and falls	1.125
Diagnostic Code: Symptoms and signs involving cognition, perception, emotional state and behavior, Symptoms and signs involving speech and voice	2.133
Diagnostic Code: Symptoms and signs involving the circulatory and respiratory systems	1.033
Diagnostic Code: Symptoms and signs involving the digestive system and abdomen	1.014
Diagnostic Code: Symptoms and signs involving the skin and subcutaneous tissue, Symptoms and signs involving the nervous and musculoskeletal systems	1.024
Diagnostic Code: Transitory endocrine and metabolic disorders specific to newborn, Digestive system disorders of newborn	0.916
Diagnostic Code: Visual disturbances and blindness, Other disorders of eye and adnexa, Intraoperative and postprocedural complications and disorders of eye and adnexa, not elsewhere classified	1.043
Disability (Memory): Has Disability	1.24
Disability (Memory): No Disability	0.783
Disability (Mobility): Has Disability	0.896
Disability (Mobility): No Disability	1.084
Disability (None): Has Disability	0.956
Disability (None): No Disability	1.016
Disability (Sensory): Has Disability	0.91

Disability (Sensory): No Disability	1.068
Ethnicity: African	1.239
Ethnicity: Any Other Asian	0.917
Ethnicity: Any Other Black	2.119
Ethnicity: Any Other Mixed	1.396
Ethnicity: Any Other White	1.377
Ethnicity: Asian or Asian British	0.583
Ethnicity: Bangladeshi	0.551
Ethnicity: Black, African, Caribbean or Black British	0.607
Ethnicity: Caribbean	0.693
Ethnicity: Chinese or Chinese British	1.153
Ethnicity: Gypsy/Gypsy Roma	0.915
Ethnicity: Indian	2.153
Ethnicity: Information not Obtained	1.609
Ethnicity: Information Refused	1.179
Ethnicity: Mixed Ethnic Groups	0.642
Ethnicity: Other Ethnic Group	0.729
Ethnicity: Pakistani	0.589
Ethnicity: Traveller	0.431
Ethnicity: White	1.226
Ethnicity: White - British	1.459
Ethnicity: White - Irish	1.656
Ethnicity: White and Asian	1.257
Ethnicity: White and Black African	0.627
Ethnicity: White and Black Caribbean	1.299
Free School Meal Status: Eligible	0.955
Free School Meal Status: Not Eligible	1.058
Free School Meal Status: Unknown	0.961

Gender: Female	0.847
Gestation Age: Continuous Value	0.972
Gestation Age: Unknown	1.236
Health Surveillance Checks Status: Not Receiving Checks	1.319
Health Surveillance Checks Status: Receiving Checks	0.85
Health Surveillance Checks Status: Unknown	0.867
Immunisation Status: Immunised	1.251
Immunisation Status: Not Immunised	0.912
Immunisation Status: Unknown	0.851
Labour Onset: Caesarean Section	0.955
Labour Onset: Medical Induction	1.184
Labour Onset: Onset Not Known	0.664
Labour Onset: Spontaneous	0.977
Labour Onset: Surgical Induction (amniotomy)	1.285
Labour Onset: Unknown	1.03
Looked After Child Status: Looked After	0.724
Looked After Child Status: Not Looked After	1.341
Maternal Smoking: 0–9 cigarettes per day	1.172
Maternal Smoking: 10–19 cigarettes per day	1.222
Maternal Smoking: 20–29 cigarettes per day	0.59
Maternal Smoking: Gave up during pregnancy	0.917
Maternal Smoking: Non-smoker	1.203
Maternal Smoking: Not Known	0.763
Maternal Smoking: Unknown	1.366
Operation Code: Approach through abdominal cavity, Approach to organ through artificial opening into gastrointestinal tract, Approach to organ through other opening, Approach to organ under image control, Harvest of nerve, Harvest of random pattern flap of skin from limb, Harvest of random pattern flap of skin from other site, Harvest of axial pattern flap of skin, Harvest of skin for graft, Harvest of flap of skin and fascia	0.789

Operation Code: Correction of congenital deformity of forearm, Correction of congenital deformity of hand, Correction of congenital deformity of hip, Correction of congenital deformity of leg, Primary correction of congenital deformity of foot, Other correction of congenital deformity of foot, Correction of minor congenital deformity of foot, Intermittent infusion of therapeutic substance, Continuous Infusion of therapeutic substance	1.116
Operation Code: Diagnostic echocardiography, Diagnostic imaging procedures, Neuropsychology tests, Nuclear medicine haematological tests, Diagnostic audiology, Breath tests, Diagnostic testing of genitourinary system, Diagnostic application tests on skin, Other diagnostic tests on skin, Diagnostic endocrinology	0.97
Operation Code: Diagnostic imaging of whole body, Diagnostic imaging of mouth, Diagnostic imaging of central nervous system, Diagnostic imaging of face and neck, Diagnostic imaging of chest, Diagnostic imaging of abdomen, Diagnostic imaging of pelvis	0.709
Operation Code: Early operations NOC, Late operations NOC, Facilitating operations NOC, Minimal access to thoracic cavity, Minimal access to abdominal cavity, Minimal access to other body cavity, Arteriotomy approach to organ under image control, Approach to organ through artery	0.89
Operation Code: Exenteration of mastoid air cells, Other operations on mastoid, Attachment of bone anchored hearing prosthesis, Repair of eardrum, Drainage of middle ear, Reconstruction of ossicular chain, Other operations on ossicle of ear, Extirpation of lesion of middle ear	1.46
Operation Code: General anaesthetic, Spinal anaesthetic, Local anaesthetic, Other anaesthetic, Y89 Brachytherapy	1.049
Operation Code: Incision of pylorus, Other operations on pylorus, Other fibreoptic endoscopic extirpation of lesion of upper gastrointestinal tract, fibreoptic endoscopic extirpation of lesion of upper gastrointestinal tract, Other therapeutic fibreoptic endoscopic operations on upper gastrointestinal tract, Diagnostic fibreoptic endoscopic examination of upper gastrointestinal tract, Therapeutic fibreoptic endoscopic operations on upper gastrointestinal tract, Intubation of stomach, Other operations on stomach, Excision of duodenum.	1.225
Operation Code: Injection of therapeutic substance, Injection of radiocontrast material, Exchange blood transfusion, Other blood transfusion, Other intravenous transfusion, Other intravenous injection, Blood withdrawal, Intramuscular injection, Subcutaneous injection, Other route of administration of therapeutic substance	0.749
Operation Code: Introduction of other inert substance into subcutaneous tissue, Introduction of destructive substance into subcutaneous tissue, Introduction of therapeutic substance into subcutaneous tissue, Introduction of substance into skin, Exploration of burnt skin of head or neck, Exploration of burnt skin of other site, Exploration of other skin of head or neck, Exploration of other skin of other site, Larvae therapy of skin, Leech therapy of skin	1.059

Operation Code: Leg region, Other vein of upper body, Other region of body, Other veins of pelvis, Laterality of operation, Other branch of thoracic aorta, Other lateral branch of abdominal aorta, Other terminal branch of aorta, Other veins of lower limb, Intervertebral disc	0.893
Operation Code: Operations on adenoid, Repair of pharynx, Other open operations on pharynx, Therapeutic endoscopic operations on pharynx, Diagnostic endoscopic examination of pharynx, Other operations on pharynx, Operations on cricopharyngeus muscle, Excision of larynx	1.04
Operation Code: Other bone of foot, Joint of shoulder girdle or arm, Joint of wrist or hand, Joint of finger, Joint of pelvis or upper leg, Joint of lower leg or tarsus, Other joint of foot, Other part of musculoskeletal system, Respiratory tract, Arm region	0.756
Operation Code: Other breech delivery, Forceps cephalic delivery, Vacuum delivery, Cephalic vaginal delivery with abnormal presentation of head at delivery without instrument, Normal delivery, Other methods of delivery, Other operations to facilitate delivery, Instrumental removal of products of conception from delivered uterus, Manual removal of products of conception from delivered uterus	0.728
Operation Code: Other closure of skin, Suture of skin of head or neck, Suture of skin of other site, Removal of repair material from skin, Removal of other inorganic substance from skin, Removal of other substance from skin, Opening of skin, Insertion of skin expander into subcutaneous tissue, Attention to skin expander in subcutaneous tissue	0.912
Operation Code: Other non-operations, External beam radiotherapy, Support for preparation for radiotherapy, Gallium-67 imaging, Radiopharmaceutical imaging, Gestational age, In vitro fertilisation, Radiology with contrast, Y98 Radiology procedures, Y99 Donor status	1.065
Operation Code: Other operations on amniotic cavity, Other therapeutic percutaneous operations on fetus, Operations on gravid uterus, Other operations on fetus, Surgical induction of labour, Other induction of labour, Elective caesarean delivery, Other caesarean delivery, Breech extraction delivery	0.905
Operation Code: Other operations on meninges of spinal cord, Therapeutic epidural injection, Drainage of spinal canal, Therapeutic spinal puncture, Diagnostic spinal puncture, Operations on spinal nerve root, excision of peripheral nerve	0.923
Operation Code: Other repair of palate, Other operations on palate, Excision of tonsil, Other operations on tonsil, Extirpation of lesion of other part of mouth, Reconstruction of other part of mouth	0.966
Operation Code: Other vascular tissue, Upper urinary tract, Lower urinary tract, Male genital organ, Vagina, Uterus, Other female genital tract, Skin of face, Skin of other part of head or neck, Skin of trunk	1.293
Operation Code: Outer ear, Other part of ear, Nose, Nasal sinus, Other respiratory tract, Mouth, Salivary apparatus, Upper digestive tract, Large intestine, Other part of bowel	0.774

Operation Code: Primary open reduction of fracture of bone and extramedullary fixation, Primary open reduction of intra-articular fracture of bone, Other primary open reduction of fracture of bone, Secondary open reduction of fracture of bone, Closed reduction of fracture of bone and internal fixation, Closed reduction of fracture of bone and external fixation, Other closed reduction of fracture of bone, Fixation of epiphysis, Other internal fixation of bone, Skeletal traction of bone	1.026
Operation Code: Radius, Ulna, Other bone of arm or wrist, Other bone of hand, Rib cage, Bone of pelvis, Femur, Tibia, Bone of tarsus	0.77
Operation Code: Simple extraction of tooth, Preprosthetic oral surgery, Surgery on apex of tooth, Restoration of tooth, Orthodontic operations, Other orthodontic operations, Other operations on tooth, Operations on teeth using dental crown or bridge, Excision of dental lesion of jaw,	0.932
Operation Code: Skin of other site, Nail, Chest wall, Abdominal wall, Muscle of shoulder or upper arm, Muscle of forearm, Muscle of hand, Muscle of hip or thigh, Muscle of lower leg, Muscle of foot	0.973
Operation Code: Ventilation support, Oxygen therapy support, Other respiratory support	0.525
Parenting Capacity (Domestic Abuse): Abuse	1.008
Parenting Capacity (Domestic Abuse): No Abuse	1.067
Parenting Capacity (Domestic Abuse): Unknown	0.903
Parenting Capacity (Learning Disabilities): Learning Disabilities	1.752
Parenting Capacity (Learning Disabilities): No Learning Disabilities	1.312
Parenting Capacity (Learning Disabilities): Unknown	0.423
Parenting Capacity (Mental Health): Mental Health Issues	1.138
Parenting Capacity (Mental Health): No Mental Health Issues	0.681
Parenting Capacity (Mental Health): Unknown	1.252
Parenting Capacity (Physical Health): No Physical Health Issues	1.4
Parenting Capacity (Physical Health): Physical Health Issues	1.935
Parenting Capacity (Physical Health): Unknown	0.359
Parenting Capacity (Substance Misuse): No Substance Misuse	1.212
Parenting Capacity (Substance Misuse): Substance Misuse	0.923
Parenting Capacity (Substance Misuse): Unknown	0.868
School Exclusion Category: Fixed Term Exclusion	1.302
School Exclusion Category: No Exclusion	0.672

School Exclusion Category: Permanent Exclusion	1.109
Substance Misuse: Misusing Substances	3.591
Substance Misuse: Not Misusing Substances	0.82
Substance Misuse: Unknown	0.33
Urban/Rural Status: 1 (Working with SAIL to identify exact meaning of this)	0.84
Urban/Rural Status: 2 (Working with SAIL to identify exact meaning of this)	1.352
Urban/Rural Status: 3 (Working with SAIL to identify exact meaning of this)	1
Urban/Rural Status: 4 (Working with SAIL to identify exact meaning of this)	1.135
Urban/Rural Status: 5 (Working with SAIL to identify exact meaning of this)	1.182
Urban/Rural Status: 6 (Working with SAIL to identify exact meaning of this)	0.588
Urban/Rural Status: 7 (Working with SAIL to identify exact meaning of this)	1.183
Urban/Rural Status: 8 (Working with SAIL to identify exact meaning of this)	0.753
Urban/Rural Status: Unknown	1.217
Welsh Index of Multiple Deprivation: Continuous Value	1.075
Welsh Index of Multiple Deprivation: Unknown	0.862
Youth Offending Status: Not Offender	0.636
Youth Offending Status: Offender	1.122
Youth Offending Status: Unknown	1.361

*Odds ratios for each risk factor are calculated by exponentiating the coefficient (logit value). Values greater than one indicate a greater odds of association between the risk factor and mental health problems, while values less than one indicate a lower odds of association between the risk factor and mental health problems.

What Works *for* Children's Social Care

EARLY INTERVENTION FOUNDATION

Coming together as What Works for Early Intervention & Children's Social Care

CONTACT

info@wweicsc.org.uk
@whatworksCSC
whatworks-csc.org.uk