What works to protect disadvantaged children and their families: a linked routine data approach

Ms Amrita Bandyopadhyay

Submitted to Swansea University in fulfilment of the requirements for the Degree of Doctor of Philosophy

Swansea University

2025

Table of Contents

Acknowledgements

“To love. To be loved. To never forget your own insignificance. To never get used to the unspeakable violence and the vulgar disparity of life around you. To seek joy in the saddest places. To pursue beauty to its lair. To never simplify what is complicated or complicate what is simple. To respect strength, never power. Above all, to watch. To try and understand. To never look away. And never, never to forget.”

— Arundhati Roy

As a human being, I find these words not only inspiring but also a guiding force in my life — a torch that illuminates my path, even when I traverse the darkest tunnels.

To my beloved Ma and Baba, you are my unwavering strength, always reminding me to pursue my dreams. This achievement is for you, and I hope to continue my exploration.

I extend my deepest thanks to my supervisor, Dr. Natasha Kennedy. Your support and guidance have been invaluable throughout this process. Sincere thanks to Professor Michael Gravenor for supporting my PhD work.

To my line manager and mentor over these 12 years at the university, Prof. Sinead Brophy, I convey my heartfelt appreciation for being a constant source of inspiration. Your encouragement has pushed me to expand the boundaries of my research and think critically. I am deeply thankful for the time and effort you have invested in my development.

I would also like to express my gratitude to the team members of The National Centre for Population Health & Wellbeing Research. Your collaboration and support have enriched my experience and contributed significantly to my work.

Additionally, I would like to thank SAIL Databank for facilitating my research efforts, providing the resources necessary for my academic pursuits, this work would not have been possible without the brilliant support of the SAIL team.

I dedicate this thesis to all those, including my closest ones, who have faced vulnerabilities in their early years and continue to fight with resilience to overcome challenges. Your strength inspires me every day. And finally, to Abhishek – today, as I reach a significant milestone of my academic journey, I am thankful for the fearlessness and strength we have developed together, and I promise to carry that spirit forward.

Thank you all for being an integral part of this journey.

Declarations and statements

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed..  .....

Date.....24 March 2026.....

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed....  .....

Date.....24 March 2026.....

I understand that the electronic version will be deposited by the repository administrator in Cronfa, the Swansea University institutional repository. The bibliographic metadata and abstract will be made immediately available. The full-text version of the thesis will be published online.

Signed....  .....

Date.....24 March 2026.....

The University's ethical procedures have been followed and, where appropriate, that ethical approval has been granted.

Signed....  .....

Date.....24 March 2026.....

List of abbreviations

ADDE	Annual District Death Extract
ADHD	Attention Deficit Hyperactive Disorder
ALF	Anonymised Linking Field
CP	Conduct Problem
DA	Domestic Abuse
DT	Decision Tree
ED	Emergency Department
EDDS	Emergency Department Dataset
EHRs	Electronic Health Records
LBW	Low Birth Weight
LR	Logistic Regression
MCSD	Millennium Cohort Study Dataset
MID	Maternity Indicators Dataset
MLR	Multivariable Logistic Regression
NCCHD	National Community Child Health Database
ONS	Office for National Statistics
PEDW	Patient Episode Dataset for Wales
PPN	Public Protection Notification
SAIL	Secure Anonymised Information Linkage
SDQ	Strengths and Difficulties Questionnaire
SMDS	Substance Misuse Dataset
WDSD	Welsh Demographic Service Dataset
WECC	Wales Electronic Cohort for Children
WLGP	Welsh Longitudinal General Practice Dataset

Summary of published works

Publication Details	Location
Bandyopadhyay, A., Jones, H., Parker, M., Marchant, E., Evans, J., Todd, C., Rahman, M. A., Healy, J., Win, T. L., Rowe, B., Moore, S., Jones, A., & Brophy, S. (2023). Weighting of risk factors for low birth weight: A linked routine data cohort study in Wales, UK. BMJ Open, 13(2), e063836. https://doi.org/10.1136/bmjopen-2022-063836	Chapter 2
Bandyopadhyay, A., Marchant, E., Jones, H., Parker, M., Evans, J., & Brophy, S. (2023). Factors associated with low school readiness, a linked health and education data study in Wales, UK. PLOS ONE, 18(12), e0273596. https://doi.org/10.1371/journal.pone.0273596	Chapter 3
Bandyopadhyay, A., Whiffen, T., Fry, R., & Brophy, S. (2023). How does the local area deprivation influence life chances for children in poverty in Wales: A record linkage cohort study. SSM - Population Health, 22, 101370. https://doi.org/10.1016/j.ssmph.2023.101370	Chapter 4
Kennedy, N., Win, T. L., **Bandyopadhyay, A.,** Kennedy, J., Rowe, B., McNerney, C., Evans, J., Hughes, K., Bellis, M. A., Jones, A., Harrington, K., Moore, S., & Brophy, S. (2023). Insights from linking police domestic abuse data and health data in South Wales, UK: A linked routine data analysis using decision tree classification. The Lancet Public Health, 8(8), e629–e638. https://doi.org/10.1016/S2468-2667(23)00126-3	Chapter 5
Bandyopadhyay, A., Brophy, S., Akbari, A., Demmler, J., Kennedy, J., Paranjothy, S., Lyons, R., & Moore, S. (2022). Health and household environment factors linked with early alcohol use in adolescence: A record-linked, data-driven, longitudinal cohort study. International Journal of Population Data Science, 7(1), Article 1. https://doi.org/10.23889/ijpds.v7i1.1717	Chapter 6
Bandyopadhyay, A., Tingay, K., Akbari, A., Griffiths, L., Bedford, H., Cortina-Borja, M., Walton, S., Dezateux, C., Lyons, R. A., & Brophy, S. (2019). Behavioural difficulties in early childhood and risk of adolescent injury. Archives of Disease in Childhood. https://doi.org/10.1136/archdischild-2019-317271	Chapter 7
Fleming, M., **Bandyopadhyay, A.,** McLay, J. S., Clark, D., King, A., Mackay, D. F., Lyons, R. A., Sayal, K., Brophy, S., & Pell, J. P. (2022). Age within schoolyear and attention-deficit hyperactivity disorder in Scotland and Wales. BMC Public Health, 22(1), 1070. https://doi.org/10.1186/s12889-022-13453-w	Chapter 8
Tingay, K., **Bandyopadhyay, A.,** Griffiths, L., Akbari, A., Brophy, S., Bedford, H., Cortina-Borja, M., Steaks, E., Walton, S., Fitzsimons, E., Dezateux, C., & Lyons, R. (2019). Record linkage to enhance consented cohort and routinely collected health data from a UK birth cohort. International Journal of Population Data Science, 4. https://doi.org/10.23889/ijpds.v4i1.579	Supplementary paper 01 (introduced in chapter1)
James, M., Rasheed, M., **Bandyopadhyay, A.,** Mannello, M., Marchant, E., & Brophy, S. (2022). The Effect COVID Has Had on the	Supplementary paper 02

Wants and Needs of Children in Terms of Play: Text Mining the Qualitative Response of the Happen Primary School Survey with 20,000 Children in Wales, UK between 2016 and 2021. International Journal of Environmental Research and Public Health, 19(19), Article 19. https://doi.org/10.3390/ijerph191912687	(introduced in chapter1)

Abstract

Aim: This thesis conducts a data-driven, population-level investigation into risk factors of early-life vulnerabilities using linked routine administrative data, integrated and harmonised with health, education and socio-economic records.

Method: The primary areas of vulnerability examined in this thesis include low birth weight, low school readiness, living in deprived areas, exposure to domestic abuse, early alcohol use, injury risk and mental health challenges. Data-driven models using advanced statistical methods (logistic regression, negative binomial regression, Cox hazard regression) and machine learning techniques (feature selection and decision trees) are employed to identify significant risk factors and their association with vulnerabilities. The Wales Electronic Cohort for Children Phase 4 has been established through this research, compiling health, education and social care data of children born or growing up in Wales.

Results: Consistent risk factors for low birth weight, low school readiness or poor academic outcomes include children living in deprivation, and poor maternal mental and physical health. Lifestyle issues such as maternal smoking, clinically significant alcohol use and substance abuse within families further exacerbate these vulnerabilities. Results reveal that children at risk of adverse outcomes, including early alcohol use and domestic abuse exposure, have fewer routine primary care contacts and more frequent emergency healthcare interactions, indicating neglect and challenging family circumstances for these children.

Conclusion: The findings demonstrate that data-driven methods can identify the signs of neglect and the associated vulnerable population from linked routine data early on in their life. This research has led to nine published papers, contributing to a strong evidence base for policies and practices aimed at improving the life chances of disadvantaged children and shaping their life trajectories.

Chapter1: Introduction

Background

The existing literature contains a significant body of work on the concept of 'vulnerability' (1). People with a greater risk of exposure to adversity, and their capacity to assess and mitigate it, are linked with conditions of vulnerability (2). In early life, individuals are dependent on the care (parental, family-level, societal) provided to them for survival and development; thus, a lack of care and support exacerbates their early-life vulnerability (3). This makes early-life vulnerability a multifaceted issue that is influenced by several factors, including health, socio-economic status and environmental conditions. These vulnerabilities often have a detrimental impact on children, potentially causing irreversible damage in areas such as physical health, emotional wellbeing, cognitive development and social skills. The early-life represents a critical developmental period, described as 'brain writing' (4), where experiences shape a child's future. Addressing these vulnerabilities and implementing mitigation strategies is crucial for preventing the long-term adverse consequences (5) and improving life trajectories. Investing in the mitigation of early-life's vulnerabilities can reduce the long-term social burden (6) while empowering marginalised and low-income families by providing them with necessary support and access to resources (7). This is critical for building a more just and equitable society. Early interventions not only improve developmental outcomes for disadvantaged children but also enhance parenting practices, leading to better overall outcomes.

In this thesis, I have conducted a data-driven exploration of linked routine administrative data to uncover risk factors associated with early-life vulnerabilities in children. The vulnerability indicators studied here are developed using linked routine data, and they include low birth weight (defined as birth weight < 2500 g), low school readiness (failure to succeed in the foundation phase), living in deprived areas (the most deprived areas as measured by the Multiple Deprivation Indicator), exposure to domestic abuse (indicated by police reports), early alcohol use (self-reported alcohol use or alcohol-related health conditions), injury risk (records of severe injury in routine healthcare data), mental health challenges, and the risk of over-diagnosis (as reported in primary care data). The risk of developing any of the above-mentioned vulnerable conditions often leads to severe adverse health and wellbeing outcomes and contributes to the public health burden. Addressing the contributing risk factors of these vulnerabilities has been a priority for the Welsh Government and Public Health Wales (PHW). This research is part of the 'Vulnerability Profiling Programme' with PHW Rhondda Cynon Taf and aligns with the early-life's priorities of the Welsh Government.

Addressing these vulnerabilities requires a holistic understanding of their determinants and the complex interplay between various influences. Data science, particularly through the use of linked routine administrative data, builds a unique framework to

analyse these factors longitudinally, uncover patterns, generate actionable insights and reveal the stories behind the data (8,9). This thesis integrates linked administrative data with advanced analytical techniques to explore childhood disadvantage. This research accumulates routinely collected administrative data from diverse sources, including health, education and social care, to identify modifiable risk factors and propose targeted interventions. The findings contribute to a growing body of evidence (10,11) supporting data-driven approaches in health and social care research, highlighting the necessity of early identification and intervention in mitigating childhood vulnerabilities. By utilising linked administrative data, this research not only enhances the existing understanding of childhood disadvantage but also equips policymakers and practitioners with evidence-based strategies to drive meaningful change (12–14).

This thesis is presented as a PhD by publication and includes nine peer-reviewed papers that collectively address key aspects of childhood vulnerability. Data science methods are utilised as tools to delve into the modifiable risk factors associated with these vulnerabilities. This thesis comprises of nine chapters. Chapters two to eight focus on seven published journal papers which form the main body of this thesis. Each chapter discusses the area of vulnerability addressed in the respective paper, the types of data used to build the research framework, the application of data science methods, including data linking, processing, harmonisation, for analysing the data. Each chapter also highlights the implications of the work and outlines my contributions to the research papers. Two other published papers are included in the Appendix, to highlight their contribution and relevance in the field of data science in health and social care research.

Methodological approaches

The identification of vulnerabilities using data science methods requires a multidisciplinary approach, incorporating knowledge of data processing and administration, statistics, advanced data modelling techniques and domain expertise (15).

Secure anonymised information linkage databank

The research was conducted using the Secure Anonymised Information Linkage (SAIL) Databank platform, located within the Medical School at Swansea University (16). SAIL is a rich and trusted repository of data, enabling comprehensive population-level studies while ensuring data privacy. Using double-encrypted anonymised person-level identifiers, known as Anonymised Linking Fields (ALF), SAIL removes identifiable and disclosive information from the data it holds. This process facilitates the longitudinal data linkage across diverse datasets while prioritising data privacy. Currently, SAIL holds over 10 billion anonymised, person-based data records, making it a significant repository of health and administrative data, supported by robust data governance practices. The integration of these extensive datasets supports a wide range of research initiatives

aimed at addressing public health and social issues. Ultimately, it contributes to improving health and social care outcomes and inform evidence-based policy decision making (17–20).

Secondary use of administrative data

In this thesis, all the accumulated research work leverages the secondary utilisation of administrative data. The large volume of data collected from numerous sources at different time points offers immense potential for secondary analysis, supporting health and social care research and decision-making policy (21). Life-course research elucidates the impact of social, economic and environmental influences on individuals' lives (22), focusing on their experiences and transitions of individuals throughout life. The secondary use of administrative data in life-course research involves repurposing datasets that were originally collected for other purposes. This approach reduces the time and cost of data collection while enabling the analysis of long-term trends and patterns. Such longitudinal data helps monitor changes in individuals' lives, identify causal relationships and better understand the dynamics of life transitions (23–25).

WECC phase 4

By using existing administrative datasets from the SAIL Databank across multiple domains, including health, education, social care, demographic and area-level information, and linking them anonymously using ALF, my research has built a platform known as the Wales Electronic Cohort for Children (WECC Phase 4). The previous three stages contained separate cohorts of children in Wales focussing on specific child health-related research questions; however, in Phase 4, I have developed a framework rather than just a single cohort of children associated with early-life vulnerability. WECC Phase 4 supports the longitudinal follow-up of a nationally representative cohort, facilitating the exploration of risk factors associated with the vulnerabilities of disadvantaged children during their early-life stages. Thus, WECC Phase 4 offers a more comprehensive and flexible research framework than the previous cohorts and this significantly improves data utilisation. This innovative platform enables me and other researchers at Swansea University to conduct comprehensive analyses of vulnerabilities. Since this data is available in the SAIL Databank, it can also be accessed by external research groups upon request, opening opportunities for future collaboration. The WECC Phase 4 platform has already secured funding for further studies. WECC Phase 4 has utilised the datasets mentioned in the table 1 (16):

Table 1: List of datasets from SAIL to build WECC Phase 4

Dataset Name	Coverage	Summary
Emergency Department Dataset (EDDS)	2009-2024	Daily version of Emergency Department Dataset.
Patient Episode Dataset for Wales (PEDW)	1995-2024	The database contains all inpatient and day case activity undertaken in NHS Wales plus

		data on Welsh residents treated in English Trusts.
Welsh Longitudinal General Practice Dataset (WLGP) - Welsh Primary Care	2000-2021	Attendance and clinical information for all general practice interactions includes patients' symptoms, investigations, diagnoses, prescribed medication and referrals to tertiary care.
Substance Misuse Dataset (SMDS)	2014-2024	The Substance Misuse Data Set captures data relating to all individuals (clients), both young persons and adults, presenting for substance misuse treatment in Wales.
Annual District Death Extract (ADDE)	1996-2024	Register of all deaths relating to Welsh residents, including those that died out of Wales
National Community Child Health Database (NCCHD)	1989-2024	The Child Health System in Wales includes birth registration and monitoring of child health examinations and immunisations.
Welsh Demographic Service Dataset (WDSD)	1990-2024	Register of all individuals registered with a Welsh GP, includes individuals anonymised address and practice history.
Education Wales (EDUW)	2004-2021	Schools and Pupil data for Wales which covers state funded learning centres.
Millennium Cohort Study Dataset (MCSD)	2000-2013	The Millennium Cohort Study (MCS), which began in 2000, is conducted by the Centre for Longitudinal Studies (CLS).
Public Protection Notification (PPN) dataset	2015-2020	PPN DASH data form South Wales Police
Tagged Electronic Cohort Cymru		Derived dataset
Children and Family Court Advisory and Support Service (CAFCASS)	2001-2024	CAFCASS (Children and Family Court Advisory and Support Service) Wales Family Justice dataset.
Lifelong Learning Wales Record (LLWR)	2017-2020	Statistics on learners in post-16 education and training, excluding those at schools but including those at Further Education Institutions, other Work-based Learning providers and Community Learning provision collected via the Welsh Government's Lifelong
Looked After Children Adoption (LACA)	1999-2021	Information about looked after children. Includes specifics on adoption, and details of adopters including ethnicity.
Maternity Indicators Dataset (MIDS)	2014-2024	The Maternity Indicators Data Set captures data relating to the woman at initial assessment and to mother and baby (or babies) for all births. This relates to initial assessment and birth activity undertaken in Wales only.

Longitudinal data linkage

To build a holistic understanding of the risk factors associated with early-life vulnerabilities, it is important to incorporate diverse resources such as health records, socio-economic data, family and environmental factors and demographic information. Consequently, data linkage has become an integral component of this research, enabling

the creation of a comprehensive dataset for studying vulnerable populations. Linking administrative data from diverse sources offers valuable opportunities to explore the multifaceted determinants of vulnerability (26). Population-level data linkage facilitates the identification of at-risk groups at both community and family levels by combining health, socio-economic and environmental influences that contribute to early-life vulnerabilities (27). This approach supports the development and improvement of targeted intervention and support plans. Since this research focuses on vulnerabilities in early-life, incorporating longitudinal data was essential to follow the study population since birth till adolescence. Data linkage enabled longitudinal analysis (28) allowing the measurement of risk factors' effects over time.

Data harmonisation and data cleaning

The research included in this thesis has significantly relied on the harmonisation of data from diverse sources, which enhances data comparability and compatibility (29). Data harmonisation improved the overall quality of the datasets and the research, increased the study population size and enhanced the generalisability of the findings (30,31). In addition to data harmonisation, data cleaning has been an integral component of all analyses conducted under this thesis. Since the research utilised real-world, routinely collected data, data cleaning was essential to identify and eliminate discrepancies, errors and anomalies inherent in such datasets. The process involved detecting outliers, inliers and unexpected patterns, as well as resolving issues such as erroneous data entries, inconsistencies and missing data. Erroneous data were removed, inaccurate values were corrected with validated replacements and missing data were handled using appropriate imputation methods.

Analysing data for vulnerability profiling

In the context of early-life vulnerability profiling, data analysis provides critical insights into the demographics of at-risk populations, including children and their families, by examining factors such as ethnicity, gender, socio-economic status and geographic location (32–34). These insights are instrumental in developing effective, targeted intervention plans for those in need. By employing a combination of traditional statistical models and machine learning approaches, a comprehensive understanding of the risk factors associated with early vulnerabilities can be achieved. Data analysis has revealed interaction patterns between these risk factors (35,36). However, the contributing nature of these factors is often complex, involving combinations of risks leading to vulnerability, such as health factors along with socio-economic deprivation (37,38), warranting further investigation.

While focusing on traditional statistical methods, the research conducted in this thesis has primarily utilised survival and regression models to investigate the association between the risk factors and the outcome. Survival analysis is used to measure the time until the event of interest occurs (39,40). Kaplan-Meier Estimator is used to visualise the

probability of an event occurring over time, this method has been used to evaluate the effectiveness of an intervention program (41). Regression analysis explores relationships between dependent outcome variables and independent exposure variables. For an outcome variable that has a continuous value, linear regression is used by fitting a linear equation (42), whereas for a binary outcome variables, logistic regression (LR) is used (43). One study used linear regression to investigate the association between child psychosocial problems and parental stress during the pandemic, with psychosocial problems measured via the Strengths and Difficulties Questionnaire (SDQ) and Parental Stress Scale (44). Similarly, another study used linear models to identify risk factors for poverty vulnerability, where poverty vulnerability levels were recorded as continuous variables (45). LR has been widely applied to examine childhood vulnerabilities by exploring associations between risk factors and binary outcomes, such as whether a vulnerability is present or absent (46–48).

In addition to traditional statistical methods, machine learning approaches such as feature selection and classification models are rapidly emerging in the field of health and social care research to gain data-driven insights. These advanced techniques effectively analyse complex datasets, uncovering patterns and relationships that may not be immediately apparent through conventional methods. The integration of machine learning supports personalised interventions tailored to individual needs. Feature selection identifies the most relevant risk factors, reducing data dimensionality, improving computational efficiency and enhancing interpretability (49,50). Supervised machine learning methods, such as classification models, are applicable for categorical outcome variables and can determine whether a child is at risk of vulnerability. This research primarily focuses on decision tree (DT) and LR models to adopt data-driven approaches for identifying vulnerabilities.

Main findings and my contributions

Main body: Data-driven approaches to investigate early-life vulnerability

I have served as the first author on six papers, including one as a joint first author and have made significant contributions as a co-author on three additional papers included in this thesis. My involvement in these research works, has encompassed all stages of the research process, including data accumulation, preparation, analysis and writing. Each paper has been published in top-quartile journals and cited by other peer-reviewed articles.

The central theme of the research presented in this thesis focuses on data-driven approaches for studying childhood vulnerabilities since birth. It establishes a robust methodological framework that employs data science techniques to identify early-life vulnerabilities and the key risk factors contributing to them. All papers were developed

within the framework of WECC Phase 4, a platform which I have established and is dedicated to life-course research on early-life vulnerabilities.

Collectively, this body of work has deepened the understanding of childhood vulnerabilities and provides actionable insights to inform future interventions aimed at supporting at-risk populations. Each of the subsequent chapters are dedicated to discussion of individual papers included in the thesis.

Supplementary paper 1: Evidence of the strength of data science approaches

The paper titled *"Record linkage to enhance consented cohort and routinely collected health data from a UK birth cohort"* is one of my published journal articles included in this thesis as supplementary work. It highlights the importance and relevance of the methodological approach implemented in this paper. By integrating the richness of longitudinal cohort data with the extensive nature of routinely collected administrative data, this study demonstrates the potential of linked datasets in investigating childhood health vulnerabilities.

Anonymously linked data from consenting participants of the Millennium Cohort Study (MCS) in Wales and Scotland, combined with their primary and secondary healthcare records, facilitated research into key health conditions, including obesity, asthma, infections, immunisations and injuries in children. This study provides a comprehensive description of data accumulation and linkage across multiple datasets (both cohort and routine) from two countries. It also compares the demographics of the populations in Wales and Scotland, identifying both similarities and differences that informed the design of subsequent research focused on childhood health vulnerabilities. The study achieved a very high linkage matching rate (above 92%) for participants in both Wales and Scotland with their administrative health data. Additionally, the paper addresses challenges encountered during the study, such as delays in data acquisition, ensuring appropriate anonymised linkage and variations in health data between the two countries. These challenges were effectively managed by implementing harmonisation process. While linking cohort data to routine health data presents complexities, it also offers significant research benefits. The key contributions of this work can be summarised as follows:

a) expanding research capacity through the integration of wider datasets.

b) providing valuable insights into public health concerns.

c) optimising resources by utilising existing data.

d) highlighting the ethical considerations of using consented and anonymised data

e) contributing to evidence-based policy decisions.

Overall, the findings underscore the potential of linked data to inform public health policies to improve health outcomes, particularly for vulnerable children. As one of the lead researchers on this project, I was responsible for data linkage, harmonisation and analysis of the data related to injury (thoroughly described in Chapter 7). I also contributed significantly to developing this research article as second author, which was published in the *International Journal of Population Data Science* and has since been cited by thirteen other research papers, as noted in Google Scholar.

Supplementary paper 2: Future direction of the methodological approaches

While data science methods are increasingly applied in health and social care research, text mining is rapidly gaining popularity in this field (51). Although this thesis primarily focuses on structured administrative data, it also acknowledges the growing importance of unstructured free-text data in healthcare and its potential to advance health and social care research (52). To illustrate the advancements in data science methods for handling unstructured data, this thesis includes one of my published journal articles, "*The Effect COVID Has Had on the Wants and Needs of Children in Terms of Play: Text Mining the Qualitative Response of the Happen Primary School Survey with 20,000 Children in Wales, UK, between 2016 and 2021*" as supplementary work.

This paper implements text mining methods to evaluate open-ended survey responses regarding children's play and the changes in these responses over time, both pre- and post-COVID lockdown. The survey captured children's feedback on improvements in their local areas, particularly in relation to their activities and play. By employing text mining techniques, we identified common themes from the free-text data, including a) time to play; b) space to play; c) permission to play; d) recommendations; and e) wellbeing outcomes. The text mining algorithms enabled the identification of the most frequently used words in children's responses, which were then mapped to key themes.

This paper demonstrates how text mining can extract meaningful insights from unstructured data patterns, providing valuable contributions to the development of intervention strategies for children's wellbeing. I developed the text mining algorithms used in this study and significantly contributed to the manuscript for publication.

Impact of my research works

My research involved accumulating routine administrative data from various sources and establishing the WECC Phase 4 platform within the SAIL Databank. The platform facilitates research pertaining to vulnerability profiling by utilising population-level data for Wales. WECC Phase 4 has been instrumental in securing research grants that has shaped my present research and will continue to guide its future trajectory.

The papers included in this thesis address various childhood vulnerabilities since birth, aligning with one of Public Health Wales's key priorities. My findings have been shared with them as a pilot study for vulnerability profiling. This collaboration highlights the relevance of my research and contributes to ongoing efforts to address early-life vulnerabilities in the region.

Conclusion

This introductory chapter has laid the foundation for the relevance and importance of applying data science methods to investigate the risk factors associated with early-life vulnerabilities. The following chapters will focus on each vulnerability in detail, from birth to adolescence, and will explore life-course research utilising data science methods on administrative data.

Chapter 2: Weighting of risk factors for low birth weight: a linked routine data cohort study in Wales, UK

Critical summary

Background

Low birth weight (LBW) is a significant indicator of childhood vulnerability, associated with long-term health and developmental challenges. The study *'Weighting of risk factors for low birth weight: a linked routine data cohort study in Wales, UK'*, published in *BMJ Open*, explores modifiable risk factors for LBW using linked routine administrative data. By integrating routinely collected administrative datasets, the study demonstrates the potential of data science to address early-life vulnerabilities, particularly among disadvantaged populations.

Utilisation of administrative data

This study brought together numerous routinely collected administrative datasets from various sources, including health and social care, to build a nationally representative cohort of children born in Wales between 1st January 1998 and 31st December 2018, forming the foundation of the research. A literature review conducted in the early stages identified explanatory variables that establish plausible causal links to LBW and are potentially modifiable at a population level. These variables span a wide spectrum, including maternal physical health (such as diabetes, anaemia, intake of vitamin D and folic acid supplements through prescription), mental health (such as depression, antidepressant medication, anxiety, serious mental illness e.g. bipolar disorder, schizophrenia), childbirth-related factors (such as maternal age, gestational age, child's birth weight, gender and birth order of the child), maternal lifestyle (such as smoking, alcohol and other substance use) socio-economic conditions (such as deprivation), living environment (such as living area) and records of domestic violence. In this study, data on maternal physical health, mental health, lifestyle factors and records of domestic violence were collected from the pregnancy period. The pregnancy period was identified as the nine months prior to the child's birth. A dataset was built that contained all the necessary risk factors by linking numerous administrative datasets to acquire the required variables. Consequently, anonymised data linkages were performed at the population level across National Community Child Health Database (NCCHD), Maternity Indicators Dataset (MIDS), Patient Episode Dataset for Wales (PEDW), Welsh Longitudinal General Practice (WLGP), Welsh Demographic Service (WDS) and Public Protection Notification (PPN) datasets. This process involved data harmonisation and thorough cleaning to ensure consistency and accuracy, facilitating the framework for analysis of the linked data. Data harmonisation was particularly important for variables like maternal smoking, as it was available from three different sources. A cleaned and

harmonised variable for maternal smoking during pregnancy, using data from the NCCHD, MIDS and WLGP datasets, was derived. While harmonising the variable, I ensured that similar codes were used for similar categories across the datasets (like 'smoker', 'non-smoker' and 'ex-smoker'). This consistency in coding was crucial for accurate comparisons across the datasets. When it came to conflicting records, a precedence rule was applied. In cases of inconsistency or missing values in one database, I prioritised the data in the following order: MIDS, then NCCHD and finally WLGP. This prioritisation was based on the coverage, consistency and nature of data collection for each database. In the MIDS, the information was recorded by health visitors during the women's pregnancies. This approach not only enhances the robustness of the findings but also allows for a more holistic examination of various risk factors that may influence LBW.

Application of data science methods

This study employed two different statistical approaches: a traditional multivariable logistic regression (MLR) model and machine learning classification model using DTs. Both methods are capable of handling large volumes of data and revealing hidden relationships between the explanatory and outcome variables. The MLR model quantifies the odds ratios associated with various risk factors, while the DT model visually represents the most significant predictors of LBW.

The MLR is a statistical method suited for predicting binary outcomes (e.g. LBW = yes/no) based on one or more independent variables (risk factors) (53). The odds ratios estimated by the MLR model indicate the strength and direction of the relationships, helping to identify the relative importance of each factor in contributing to LBW (54). This quantification is essential for public health interventions, as it enables policymakers to prioritise resources and strategies based on the most significant risk factors. Additionally, MLR simplifies interpretation and facilitates hypothesis testing.

In contrast, the DT model effectively handles complex and non-linear relationships between independent and outcome variables. DTs provide a clear, visual representation of the decision-making process, enhancing the interpretability and communication of findings to stakeholders, including healthcare providers and policymakers (55). This model excels in identifying subgroups at higher risk for LBW by segmenting the data based on predictor values, revealing insights that may be obscured in traditional regression analyses (56).

Integrating these two approaches provides a comprehensive, data-driven understanding of LBW from the linked datasets. By applying these methods, I identified a range of modifiable risk factors, including maternal health conditions, lifestyle choices and socio-economic deprivation. The findings highlight the importance of addressing these factors

to reduce LBW prevalence and improve health outcomes for disadvantaged children and their families.

Early-life vulnerability profiling

The paper's emphasis on profiling early-life vulnerability is particularly important, as this can lead to several detrimental effects on individuals, including impaired cognitive development, disability and poor academic achievement. By identifying the risk factors associated with LBW, the study contributes to a broader understanding of the vulnerabilities faced by disadvantaged children from the outset of life. The findings suggest that interventions aimed at improving maternal health, promoting healthy lifestyles and addressing socio-economic disparities could have a significant impact on reducing LBW and its associated risks.

Moreover, the research highlights the need for targeted public health strategies that address the unique challenges faced by disadvantaged families. By profiling these vulnerabilities, the study establishes a foundation for developing tailored interventions that can effectively support at-risk populations.

Published journal paper

BMJ Open

Weighting of risk factors for low birth weight: a linked routine data cohort study in Wales, UK

Amrita Bandyopadhyay ,[1] Hope Jones ,[1] Michael Parker,[1] Emily Marchant ,[1] Julie Evans,[2] Charlotte Todd,[2] Muhammad A Rahman,[3] James Healy,[1,4] Tint Lwin Win,[1] Ben Rowe,[5] Simon Moore ,[6,7] Angela Jones,[2] Sinead Brophy [1]

To cite: Bandyopadhyay A, Jones H, Parker M, *et al*. Weighting of risk factors for low birth weight: a linked routine data cohort study in Wales, UK. *BMJ Open* 2023;**13**:e063836. doi:10.1136/bmjopen-2022-063836

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-063836).

Received 19 April 2022
Accepted 28 November 2022

For numbered affiliations see end of article.

Correspondence to
Ms Amrita Bandyopadhyay;
[redacted]

ABSTRACT

Objective Globally, 20 million children are born with a birth weight below 2500 g every year, which is considered as a low birthweight (LBW) baby. This study investigates the contribution of modifiable risk factors in a nationally representative Welsh e-cohort of children and their mothers to inform opportunities to reduce LBW prevalence.

Design A longitudinal cohort study based on anonymously linked, routinely collected multiple administrative data sets.

Participants The cohort, (N=693 377) comprising of children born between 1 January 1998 and 31 December 2018 in Wales, was selected from the National Community Child Health Database.

Outcome measures The risk factors associated with a binary LBW (outcome) variable were investigated with multivariable logistic regression (MLR) and decision tree (DT) models.

Results The MLR model showed that non-singleton children had the highest risk of LBW (adjusted OR 21.74 (95% CI 21.09 to 22.40)), followed by pregnancy interval less than 1 year (2.92 (95% CI 2.70 to 3.15)), maternal physical and mental health conditions including diabetes (2.03 (1.81 to 2.28)), anaemia (1.26 (95% CI 1.16 to 1.36)), depression (1.58 (95% CI 1.43 to 1.75)), serious mental illness (1.46 (95% CI 1.04 to 2.05)), anxiety (1.22 (95% CI 1.08 to 1.38)) and use of antidepressant medication during pregnancy (1.92 (95% CI 1.20 to 3.07)). Additional maternal risk factors include smoking (1.80 (95% CI 1.76 to 1.84)), alcohol-related hospital admission (1.60 (95% CI 1.30 to 1.97)), substance misuse (1.35 (95% CI 1.29 to 1.41)) and evidence of domestic abuse (1.98 (95% CI 1.39 to 2.81)). Living in less deprived area has lower risk of LBW (0.70 (95% CI 0.67 to 0.72)). The most important risk factors from the DT models include maternal factors such as smoking, maternal weight, substance misuse record, maternal age along with deprivation—Welsh Index of Multiple Deprivation score, pregnancy interval and birth order of the child.

Conclusion Resources to reduce the prevalence of LBW should focus on improving maternal health, reducing preterm births, increasing awareness of what is a sufficient pregnancy interval, and to provide adequate support for mothers' mental health and well-being.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This study has built an e-cohort using data-linkage across multiple routinely collected administrative data sets to investigate the risk factors of low birth weight (LBW) for the population of Wales.
⇒ The study has investigated the modifiable risk factors of LBW in a holistic framework by linking primary and secondary care physical and mental health, socio-demographic and pregnancy-related routine data including police record for a nationally representative sample.
⇒ This study undertook two different statistical approaches (regression analysis and data-driven machine learning algorithm) which is a strength of the study.
⇒ This work was unable to include any important risk factors which were not recorded in the healthcare system or any conditions which were undiagnosed hence that did not result in the system.

INTRODUCTION

The WHO defines low birth weight (LBW) as infants weighing less than 2500 g (5.5 pounds) irrespective of gestational age.[1 2] Latest figures show that each year around 53 000 live births (6.9%) are identified as LBW in the UK.[3] LBW is the result of intrauterine growth restriction (less than 10th centile of weight for sex and gestational age), prematurity (gestational age less than 37 weeks) or a combination of both.[4] LBW can impair the baby's cognitive development and lead to developmental disabilities and poor academic achievement.[5] Furthermore, LBW significantly increases the risk of perinatal and neonatal mortality and longstanding morbidity in early and later life.[6] While there has been a reduction in mortality among preterm infants in the last two decades, the incidence of preterm birth has increased in many developed countries.[6–8] The increase is also associated with preterm delivery of multiple pregnancies,

with medically indicated preterm birth 10 times more likely in multiple pregnancies than singleton births.[9] To address the global burden of LBW, the 65th World Health Assembly Resolution 65.6 endorsed a comprehensive implementation plan to achieve a 30% reduction in LBW by 2025.[1] A study conducted on the birth data from 148 countries of 195 United Nations' member states indicated that there had been a 2.9% reduction in the LBW prevalence in 2015, compared with 2000 worldwide. However, there has not been any change in the LBW prevalence in high-income regions (including Europe) and the progress is slower than required to meet the WHO LBW target by 2025.[10]

Existing research has found factors linked with mothers, such as age, high deprivation and low academic qualification, are associated with increased odds of LBW.[11 12] Modifiable risk factors for LBW include interpregnancy interval,[13] maternal physical[14–17] and mental health[18 19] and environmental exposures during pregnancy.[20] Studies have also shown numerous health behaviours such as smoking,[21 22] alcohol intake (in which there is a dose-response relationship with LBW)[23] and/or illicit drug use[24] during pregnancy are modifiable risk factors of LBW. Indirect (negative maternal behaviours, inadequate nutrition or prenatal care and increased stress) or direct (physical assault, sexual trauma) experience of intimate partner abuse during pregnancy can lead to adverse infant outcomes including LBW.[25 26]

It is important to gain an understanding of these risk factors, particularly modifiable risk factors, so that resources and interventions can be scheduled effectively. Moreover, the wide range of risk factors cannot be addressed in isolation. Most of the risk factors that are strongly independently associated with LBW are correlated. This study aimed to understand the contributions of risk factors to the burden of LBW for the population of Wales, using traditional statistical methods and supervised machine learning models.

METHOD

Participants and linkage

The linked data cohort (N=693 377) comprised of children born in Wales between 1 January 1998 and 31 December 2018. The study population was identified in the National Community Child Health Database (NCCHD), which is a local Child Health System database held by the National Health Service. The participants were linked to the Wales-wide administrative register, the Wales Demographic Service (WDS) dataset. Linkage was undertaken using an anonymised encrypted linkage key, the anonymised linking field, in the Secure Anonymised Information Linkage (SAIL) Databank.[27] WDS provided the anonymised residential linking fields, which is an encrypted residential address and its corresponding lower super output area (LSOA, small geographical areas with a population of approximately 1500) when the child was born. LSOA was linked with the Welsh Index of Multiple Deprivation (WIMD) 2014, which is a measure of relative deprivation. The participants flow diagram is displayed in figure 1.

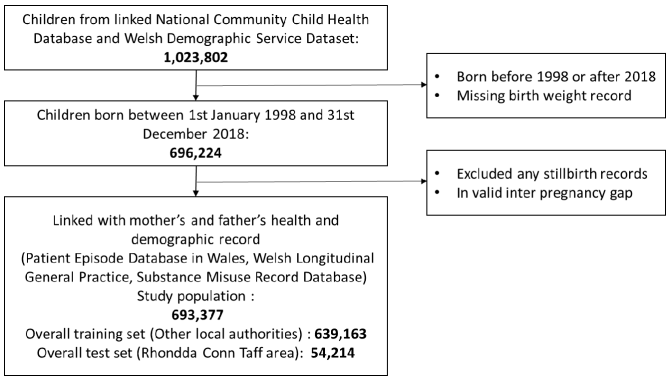


Figure 1 Participants flow diagram.

Explanatory variables

A literature review was conducted at the beginning of the study to identify the explanatory variables associated with LBW. A study by Johnson *et al* was identified[3] and this provided the framework on which the current study was developed. The literature review selected:

1. Any published systematic reviews since 2013 which focused on risk factors identified in Johnson *et al*.
2. Any published systematic reviews since 2010 for all additional risk factors not identified in Johnson *et al*.

This study therefore considered a wide range of explanatory and confounding variables that have a plausible causal link to LBW and are potentially modifiable at a population level. The literature review to select the explanatory variables has been described in a online supplemental document Supplementary document. In the current study, modifiable risk factors identified from the literature have been derived from routinely collected electronic datasets to build a Welsh e-cohort of the children. The maternal variables related to a childbirth (maternal age, gestational age, child's birth weight, gender and birth order of the child) were obtained from NCCHD and maternal indicator database (MID). The variables for maternal physical (such as diabetes, anaemia, intake of vitamin D and folic acid supplement through prescription) and mental (depression, antidepressant medication, anxiety, serious mental illness such as bipolar disorder, schizophrenia) health during pregnancy were obtained from primary care Welsh Longitudinal General Practice (WLGP) and hospital admissions dataset known as the Patient Episode Database in Wales (PEDW). The record of physical assault linked with mothers during pregnancy was obtained from PEDW. The substance misuse database provided the information on individuals receiving treatment for alcohol and other substance misuse in Wales. Mothers' who were presenting in this database during pregnancy were considered in the study. Area type (urban/rural) and local authority (LA) under which they lived during the pregnancy and their overall and physical environment quantified in the WIMD were

included in this study. A cleaned and harmonised variable of maternal smoking during pregnancy was created based on the data obtained from NCCHD, MID and WLGP datasets. The other derived maternal variables include multiple birth flag (to distinguish between singleton and non-singleton), pregnancy interval and maternal weight. The description of the explanatory variables and their sources have been described in online supplemental table 1.

A subset of the study population (participants from Rhondda, Cynon, Taf, born between June 2016 and 2018) was linked with the Public Protection Notification (PPN) dataset to investigate the impact of the PPN during pregnancy along with other existing risk factors on the risk of LBW.[28] PPN is an information sharing system, completed by police officers that compiles incidents of domestic abuse, stalking or harassment. The current study received PPN data from South Wales Police for residents of South Wales LA Rhondda, Cynon, Taf.

Outcome variable

A binary variable was created using the birth weight variable obtained from NCCHD.

- LBW=birth weight <2500.
- Not LBW=birth weight ≥2500.

Statistical analysis

It is known that gestational age is highly correlated with LBW. However, as the gestational age is only obtained at the point of birth, making it a non-modifiable risk factor, this study has not considered it as a predictor variable. The models were stratified by the multiple birth as this is one of the main predictors of LBW. The missing records in the birthweight variable were removed from the analysis. Since there was around 15% missing data in the maternal weight variable, the variable was imputed by the simple random imputation method.[29] The missing data in the other explanatory variables (less than 10%) were recorded as 'Unknown'. The birth record for stillbirth and pregnancy interval of less than 22 weeks (as that is the minimum duration for a considerable gestation period) were also not considered for the statistical analysis. Data preparation including data linkage and data cleaning for this analysis was done on SAIL DB2 SQL platform. All statistical analyses were performed in R V.4.0.3.

The statistical analysis of the current study was carried out using two statistical approaches: (a) building a holistic regression model to investigate the association between the risk factors and LBW and (b) building a predictive model using a supervised classification method. Both methods were capable of handling binary outcome variables. The models that were developed by the above-mentioned methods were built independently, however they both were informed by the same dataset. This enabled us to evaluate and validate the findings of the models and helped to gain insight on the generalisability of the findings.

Logistic regression

A multivariable logistic regression (MLR) model was developed to identify the most important risk factors associated with LBW. The MLR model was built on the overall study population (whole Wales dataset) to examine the associations between all the explanatory and outcome variables. The holistic model considering all the risk factors identified from literature review and selected or derived from routine data includes maternal physical and mental health during pregnancy, maternal smoking, alcohol and other substance misuse record, maternal age, maternal weight, pregnancy interval, living area, LA and deprivation—WIMD score. The MLR model also included the birth order of the child and the multiple birth flag. The birth order highlights the sequential birth position of the child for a mother, and it does not vary among the children who were non-singleton in the same family (please see online supplemental table 1), hence, they were considered as independent variables in the model and their association with the outcome variable was investigated in the MLR model. The importance and significance of the risk factors have been evaluated and presented with their adjusted OR (aOR) and 95% CI.

Decision tree

A supervised machine learning classifier—decision tree (DT) model was developed to build a risk profile for LBW and test its predictive performance. Classification tree—DT models were constructed using RPART (Recursive Partitioning And Regression Trees) packages in R.[30 31] The algorithm recursively partitions the data into multiple subspaces to obtain the homogeneous final subspace of predictor variables. For DT, the whole Wales data except for Rhondda, Cynon, Taf, was used to train the model and prediction performance was evaluated on a test dataset which consisted of a sample of participants from the LA of Rhondda, Cynon, Taf. This LA was chosen because it had one of the highest rates of LBW in Wales and is an area which would benefit most from an accurate prediction model.

A separate data linkage was undertaken with a subset of the study population which was linked to the mother's domestic abuse record from PPN dataset (the latter was only available for Rhonda, Cynon, Taf). Another adjusted MLR model was developed on this linked data to investigate the risk association for LBW.

Patient and public involvement

No patient involved.

RESULTS

The study population consisted of 693 377 children of which 54 214 were from Rhondda, Cynon, Taf, and 639 163 were from other LAs. The children from Rhondda, Cynon, Taf, which was later used as a test set for DT were well representative of the Welsh population (see online supplemental table 2). In the overall study

population, 51.26% were boys, 96.92% were singleton and 90.38% children were born full-term (gestational age between 37 and 42 weeks). 49.85% of the children were born as the first child in the family. Mothers of 0.48% children were admitted to hospital for diabetes and 0.09% had a general practitioner (GP) visit for diabetes, 1.27% had depression, 1.52% with anxiety and 0.02% were on antidepressant medication during pregnancy. There were 1.26% and 21.51% children whose mothers had alcohol-related substance misuse and smoking records during pregnancy, respectively. The average maternal age at birth of child and maternal weight was 28 years and 70.82 kg (after imputation), respectively, and 63.68% of them were living in densely populated urban areas. Overall, 7.1% (8.26% in test set and 7% in other LAs) of children were born as LBW.

Factors associated with LBW: MLR results

Non-singleton children were at almost 22 times higher risk of LBW than singleton children (aOR—21.74 (95% CI 21.09 to 22.40)). Mothers with diabetes-related GP visits (2.03 (95% CI 1.81 to 2.28)) and hospital admission records of anaemia (1.26 (95% CI 1.16 to 1.36)) during pregnancy were at very high risk of having LBW children. Poor mental health during pregnancy such as severe depression (1.58 (95% CI 1.43 to 1.75)), serious mental illness (1.46 (95% CI 1.04 to 2.05)), severe anxiety (1.22 (95% CI 1.08 to 1.38)) and antidepressant medications (1.92 (95% CI 1.20 to 3.07)) were risk factors for LBW. The other highly significant modifiable risk factors linked with pregnant mothers include maternal smoking (1.80 (95% CI 1.76 to 1.84)), alcohol-related hospital admissions (1.60 (95% CI 1.30 to 1.97)) and any substance misuse (alcohol/other drugs) (1.35 (95% CI 1.29 to 1.41)) during pregnancy. Higher maternal age was also associated with the risk of LBW. Though maternal age less than 19 was significantly associated with the risk of LBW in the univariable model, after adjusting all the other explanatory variables, this did not remain as a risk factor of LBW. The first child born was at higher risk of LBW than subsequent births, The odds of LBW for the second child was 0.59 (95% CI 0.57 to 0.60) compared with the first child. Mothers living in the least deprived and rural areas during pregnancy were at lower risk of having LBW children than others living in more deprived and urban areas. The statistically significant risk factors with their aOR and CI have been visualised and described in figure 2 and online supplemental table 3.

Finding from the linked PPN data model

A data set of 5854 mothers were obtained from the PPN data linkage. Those who had a PPN call during pregnancy, 18% of them had an LBW child and those who did not have a PPN call, 8.7% of them had an LBW child (see table 1). Mothers with a PPN call during pregnancy had almost two times higher risk of having LBW babies (1.98 (95% CI 1.39 to 2.81)) than mothers without PPN call after adjusting for confounding factors (see online supplemental figure 1).

Predictive DT model

Since LBW were disproportionately more prevalent in non-singleton children (5.61% singleton vs 53.91% of the non-singleton children were LBW) (online supplemental table 4), two separate predictive models using DTs were developed.

Singleton children

There were 619 458 observations in the training model. The most important risk factors selected by the DT algorithm to develop the final tree were maternal smoking, maternal weight, pregnancy interval, birth order, maternal substance misuse record (any), maternal age, deprivation—WIMD score, maternal substance misuse record (other drug) and maternal substance misuse record (alcohol). Online supplemental figure 2 depicts the final tree with the branches including the final 33 terminal nodes. For example, the model would predict an LBW baby if (a) maternal smoking is positive (eg, mum smokes during pregnancy) and (b) maternal weight less than 60 kg. The number of women in this category who had an LBW child is 73% (see terminal node 4 in online supplemental figure 2) and risk profile was found in 7% of the training model population (eg, 7% of pregnant women were smokers who weighed less than 60 kg during pregnancy).

The test data was built on the 52 583 singleton children, which is 7.82% of the total singleton children in this study. The model performance is explained in a confusion matrix with 60.54% accuracy, 60.41% sensitivity, 60.55% specificity, 9.68% positive predictive values and 95.63% negative predictive value (see tables 2,3).

Non-singleton children

There were 19 705 children in the non-singleton training subset. The variables selected to generate the tree by the DT algorithm in the importance order were pregnancy interval, birth order, maternal weight, maternal age, gender, deprivation—WIMD score, maternal smoking, living area, deprivation—WIMD (environment) score and maternal substance misuse record (any). Online supplemental figure 3 depicts the final tree with the branches including the final 29 terminal nodes. For example, the model would predict an LBW baby if (a) this is the first child or pregnancy interval is either above 10 years or less than 1 year and (b) maternal weight less than 60 kg (terminal node 4).

The test set was built on the 1631 non-singleton children, which is 7.64% of the total non-singleton children in this study. The model performance was measured as 58.74% accuracy, 68.71% sensitivity, 41.09% specificity, 67.36% positive predictive values and 42.61% negative predictive value (see tables 2,3).

DISCUSSION

Among the overall study population in Wales 7.1% was LBW between 1998 and 2018. Global trend of LBW is

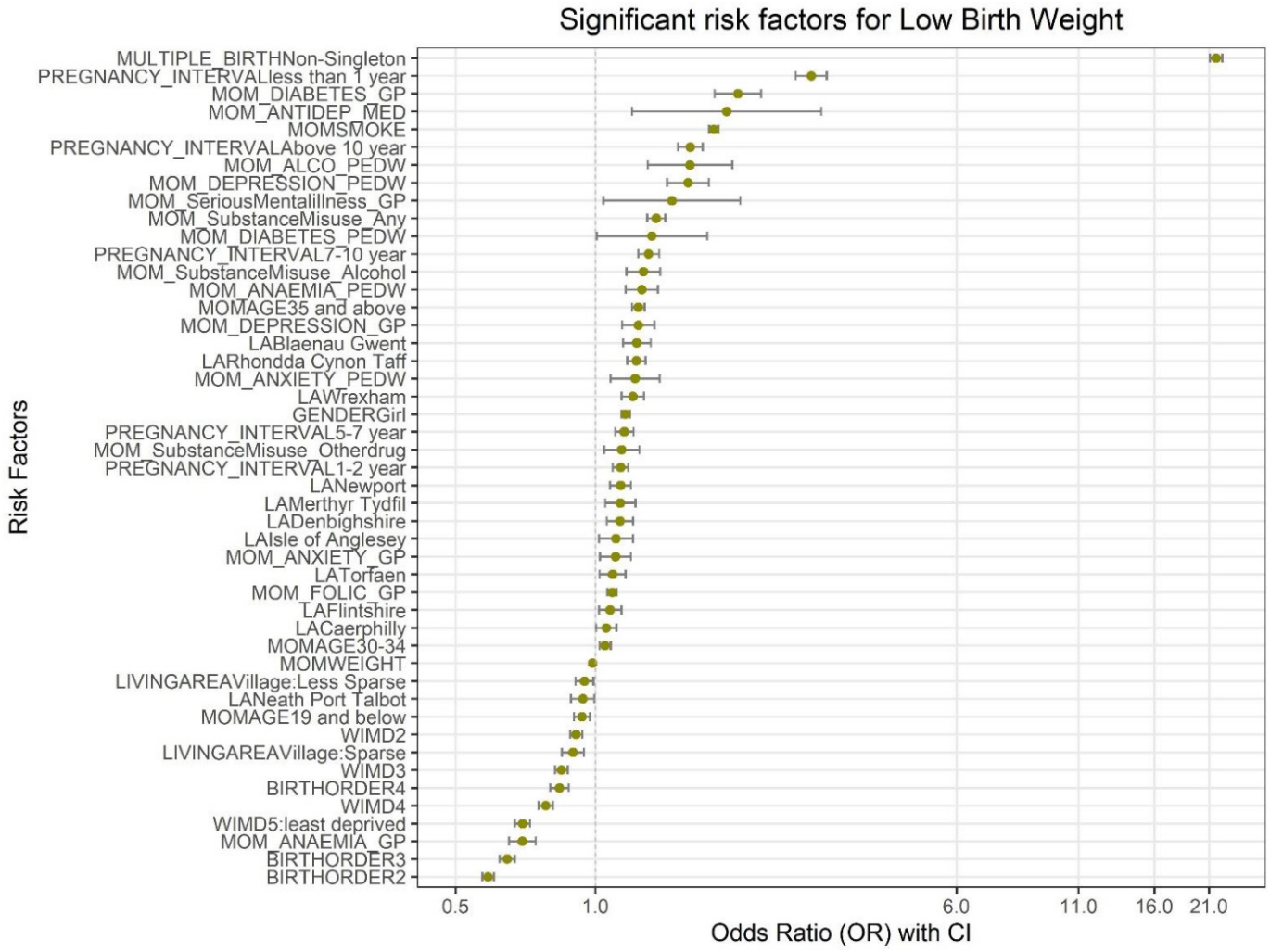


Figure 2 Significant factors associated with the risk low birth weight among the overall study population. GP, general practitioner; LA, local authority; PEDW, Patient Episode Database in Wales; WIMD, Welsh Index of Multiple Deprivation.

around 7.0% in both 2000 and 2015 for the developed regions (Europe, North America, Australia), which is consistent with our finding.[2] Findings from the Office for National Statistics state a combined English and Welsh rate of LBW of 7.0% in 2016, unchanged from 2011.[32] Our findings show that LBW is strongly associated with non-singleton pregnancy, and maternal health which includes a short pregnancy interval, non-optimal maternal body weight (eg, low, or high weight), maternal smoking, diabetes, anaemia, mental illness and living in a deprived urban area and exposed to domestic abuse during pregnancy.

Table 1 Distribution of LBW and nLBW children for the subset who were linked with mother's PPN record during pregnancy

PPN record during pregnancy	n=5854	
No		
nLBW	5074	91.3%
LBW	485	8.7%
Yes		
nLBW	241	82%
LBW	53	18%

LBW, low birth weight; nLBW, not LBW; PPN, public protection notification.

Table 2 Confusion matrix/two by two table of the decision tree (singleton and non-singleton) models

	Reference (singleton) n=52 583		Reference (non-singleton) n=1631	
Prediction	**LBW**	**nLBW**	**LBW**	**nLBW**
LBW	2077 (TP)	19 389 (FP)	716 (TP)	347 (FP)
nLBW	1361 (FN)	29 756 (TN)	326 (FN)	242 (TN)

FN, False Negative; FP, False Positive; LBW, low birth weight; nLBW, not LBW; TN, True Negative; TP, True Positive .

Table 3 Prediction model performance (n=52 583 singleton, n=1631 non-singleton from test set)

	Accuracy	Sensitivity	Specificity	Positive predictive value	Negative predictive value
DT singleton model	60.54%	60.41%	60.55%	09.68%	95.63%
DT non-singleton model	58.74%	68.71%	41.09%	67.36%	42.61%

DT, decision tree.

The findings of short and long pregnancy intervals being associated with increased odds of LBW has been reported previously.[13] However, Regan *et al* highlighted that several studies examining long interpregnancy intervals are prone to measurement error because miscarriages and abortions within this time period are difficult to capture. Hence the authors suggest that caution should be exercised when interpreting these findings.[33] Regarding the association of short-pregnancy intervals with increased odds of LBW, studies using matched controlled designs have argued that this association may be weaker than previously thought,[33 34] especially when adjusting for factors such as gestational diabetes, pre-pregnancy obesity, parity and other familial factors.[35] The current study has included diabetes and maternal weight along with pregnancy interval in the analysis. In terms of putting this evidence in context, when considering advice over pregnancy intervals, it will be important to consider all the available evidence including the impact of pregnancy interval on preterm birth and maternal outcomes.[36] Among the modifiable risk factors for LBW identified in this study, smoking during pregnancy is significantly and consistently important. A number of reviews have been carried out in the field of interventions to reduce smoking in pregnancy and this suggest that psychosocial interventions (counselling, feedback and incentives) appear to be effective at supporting women to stop smoking in pregnancy which, in turn, can reduce the proportion of babies born with LBW.[37] However, they argue that the context of the intervention needs to be given consideration and that while evidence exists for potentially effective interventions which could be piloted through delivery of programmes locally, efforts should also be directed at population wide strategies to reduce smoking uptake in young women. This may be especially important given the clear difficulties experienced by pregnant women to give up smoking.[37] With regards to our finding of maternal mental health affecting the risk of LBW, both severe depression and anxiety were associated with an increased odds of LBW in our study.[38]

The study undertook two statistical methods; (a) regression and (b) supervised classification model with the aim that the regression model would identify the risk factors with highest association/OR but not frequently observed factors at the population level for, for example, only 0.09% mothers had diabetes-related GP visit during pregnancy, and they had two times higher risk of having a LBW child (2.03 (95% CI 1.81 to 2.28)). However, the DT models consider the number of people affected by the risk factor rather than just strength of association, hence capable of identifying the factors at a population level (such as smoking, deprivation score) that can result in higher risk of LBW.

There are similarities between the findings of our DT models and existing literature using machine learning to predict LBW, for example, urban living, higher deprivation and poorer families are at higher risk of LBW.[39] The incidence of LBW in this current work is lower than another research using machine learning to predict LBW, for example, Loreto *et al* has an incidence of 13.45% in work that builds over 60 different machine learning models,[40] Ahmadi *et al* assess logistic regression and random forests in a cohort with LBW rate of 9.5%.[41] The smaller number of active cases in the dataset the more difficult it is to build a prediction model for, particularly without a set of highly associated input variables. In this study, the singleton DT model correctly predicted 60.41% of all the true positive cases. However, the low positive predictive value of 9.68% indicates that the model assigned a false positive 'LBW' classification for 89.32% cases. This model only includes singleton children and since non-singleton pregnancies are highly associated with LBW, removing this variable from the model has lessened its predictive capability. This is evidenced by the significantly improved positive predictive value (67.36%) for the non-singleton model (table 3). Previous machine learning models appear to show better prediction as they included non-singleton, gestational age (which is in terms of temporal association highly associated with LBW but occurs at the same time as the LBW can be measured) and pre-eclampsia in the third trimester. Also, the differences in the proportion of LBW cases, the variables used and the cohort sizes in various other studies alter the ability of the model, hence direct comparison of machine learning models across studies can become difficult.

The strength of this study lies in using a wide spectrum of routinely collected nationally representative administrative data sets of all births in Wales across a large time. This is a very first of its kind study in Wales and adds novelty in the research field of LBW. However, this work can only identify the more severe cases which are recorded in the healthcare system, and undiagnosed cases that did not result in the system will be missed which is a limitation of this work. Since the study was developed on the linked routine data, the limitation of the routine data was encountered in this study, for example, though

the maternal weight variable came from two different sources, data was missing for many participants which was addressed by imputation methods. Also, this study was unable to capture lifestyle factors (diet, physical activity, stress, emotional state) which can be important in determining LBW.[42 43]

The two different models (MLR and DT) used in this study have very similar findings suggesting that factors which are common and so are predictive (using DT methods) such as maternal smoking status and maternal weight could be targeted to address population-level risk of LBW. Factors which have a strong association with LBW (using regression analysis), such as a mother with diabetes or mother on antidepressants as having plausible causal link to LBW, can be addressed to reduce individual risk for that mother/child.

CONCLUSION

This study suggests that the most important factors to reduce the risk of LBW are to address multiple birth (eg, in assisted reproduction practices), addressing factors associated with preterm births (previous history of preterm birth), addressing maternal health such as reducing smoking, investment in maternal mental health, addressing substance use (alcohol/drugs), treating underlying health conditions (diabetes/anaemia) and promoting planning of pregnancy to give an adequate pregnancy interval and healthy weight of mother especially for those in deprived urban areas.

Author affiliations
[1]National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Swansea, UK
[2]Keir Hardie University Health Park, Public Health Wales, Cardiff, UK
[3]Cardiff School of Technologies, Cardiff Metropolitan University, Llandaff Campus, Cardiff, UK
[4]Office for National Statistics, Government Buildings, Cardiff Road, Duffryn, Newport, UK
[5]National Police Chiefs' Council Lead for Mental Health and Age, London, UK
[6]Violence Research Group, School of Dentistry, Cardiff University, Cardiff, UK
[7]Security, Crime, Intelligence Institute, Cardiff University, SPARK, Maindy Road, Cardiff, UK

Twitter Emily Marchant @emily_marchant and Sinead Brophy @@SineadBr

Contributors Planning—conceptualisation: SB, AJ, JE and AB. Data acquisition: The police Public Protection Notification and Maternal Indicator Database data accusation was supported by BR and JE, respectively. The other health data was available in Secure Anonymised Information Linkage, obtained through Information Governance Review Panel (IGRP) request led by SB and AB. Supervision: SB. Conduct—literature review: CT and EM. Methodology: AB and SB. Data preparation: MAR and AB. Formal analysis and investigation: AB. Additional support in analysis: JH and MP. Writing—original draft preparation: AB. Review and editing: CT, MP, JE, EM, HJ, MAR, JH, TLW, BR, SM, AJ and SB. All authors read and approved the final manuscript. SB and AB are responsible for the overall content.

Funding This work was funded by Public Health Wales (PHW), grant number (105186). This work was supported by National Institute for Health Research (NIHR), grant number (NIHR133680). This research has been carried out as part of the ADR Wales programme of work. The ADR Wales programme of work is aligned to the priority themes as identified in the Welsh Government's national strategy: Prosperity for All. ADR Wales brings together data science experts at Swansea University Medical School, staff from the Wales Institute of Social and Economic Research, Data and Methods (WISERD) at Cardiff University and specialist teams within the Welsh Government to develop new evidence which supports Prosperity for All by using the Secure Anonymised Information Linkage (SAIL) Databank at Swansea University, to link and analyse anonymised data. ADR Wales is part of the Economic and Social Research Council (part of UK Research and Innovation) funded by ADR UK (grant ES/S007393/1). This work was also supported by the National Centre for Population Health and Well-Being Research (NCPHWR) which is funded by Health and Care Research Wales. This work was supported by Health Data Research UK which receives its funding from HDR UK Ltd (NIWA1) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Wellcome Trust. This work uses data provided by patients and collected by the National Health Service as part of their care and support. This study used anonymised data held in the SAIL Databank. We would like to acknowledge all the data providers who enable SAIL to make anonymised data available for research. We acknowledge the support provided by South Wales Police. The work conducted does not represent or is it endorsed by the Office for National Statistics.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. The data have been archived in the Secure Anonymised Information Linkage Databank (https://saildatabank.com/0029)

ORCID iDs
Amrita Bandyopadhyay http://orcid.org/0000-0003-2798-4030
Hope Jones http://orcid.org/0000-0003-4312-476X
Emily Marchant http://orcid.org/0000-0002-9701-5991
Simon Moore http://orcid.org/0000-0001-5495-4705
Sinead Brophy http://orcid.org/0000-0001-7417-2858

REFERENCES

1 WHO. WHO | global nutrition targets 2025: low birth weight policy brief. n.d. Available: http://www.who.int/nutrition/publications/globaltargets2025_policybrief_lbw/en/
2 UNICEF-WHO low birthweight estimates: levels and trends 2000–2015. Available: https://www.unicef.org/reports/UNICEF-WHO-low-birthweight-estimates-2019 [Accessed 3 Mar 2022].
3 Johnson CD, Jones S, Paranjothy S. Reducing low birth weight: prioritizing action to address modifiable risk factors. *J Public Health (Oxf)* 2017;39:122–31.
4 Mohammed SG. Low birth weight in omdurman maternity hospital. *Int J Sci Res Publ* 2014;4:1–13.
5 Breslau N, Paneth NS, Lucia VC. The lingering academic deficits of low birth weight children. *Pediatrics* 2004;114:1035–40.
6 Ohlsson A, Shah P. *Determinants and prevention of low birth weight: a synopsis of the evidence*. Institute of Health Economics, 2008.
7 Heaman MI, Sprague AE, Stewart PJ. Reducing the preterm birth rate: a population health strategy. *J Obstet Gynecol Neonatal Nurs* 2001;30:20–9.

8 Yuan W, Duffner AM, Chen L, *et al*. Analysis of preterm deliveries below 35 weeks' gestation in A tertiary referral hospital in the UK. A case-control survey. *BMC Res Notes* 2010;3:119.
9 Blencowe H, Cousens S, Chou D, *et al*. Born too soon: the global epidemiology of 15 million preterm births. *Reprod Health* 2013;10 Suppl 1(Suppl 1):S2.
10 Blencowe H, Krasevec J, de Onis M, *et al*. National, regional, and worldwide estimates of low birthweight in 2015, with trends from 2000: a systematic analysis. *Lancet Glob Health* 2019;7:e849–60.
11 Shi L, Macinko J, Starfield B. Primary care, infant mortality, and low birth weight in the states of the USA. *J Epidemiol Community Health* 2004;58:374–80.
12 Silvestrin S, Silva C da, Hirakata VN, *et al*. Maternal education level and low birth weight: a meta-analysis. *J Pediatr (Rio J)* 2013;89:339–45.
13 Conde-Agudelo A, Rosas-Bermúdez A, Kafury-Goeta AC. Birth spacing and risk of adverse perinatal outcomes: a meta-analysis. *JAMA* 2006;295:1809–23.
14 Yu Z, Han S, Zhu J, *et al*. Pre-pregnancy body mass index in relation to infant birth weight and offspring overweight/obesity: a systematic review and meta-analysis. *PLOS ONE* 2013;8:e61627.
15 Daalderop LA, Wieland BV, Tomsin K, *et al*. Periodontal disease and pregnancy outcomes: overview of systematic reviews. *JDR Clin Trans Res* 2018;3:10–27.
16 Flynn CA, Helwig AL, Meurer LN. Bacterial vaginosis in pregnancy and the risk of prematurity: a meta-analysis. *J Fam Pract* 1999;48:885–92.
17 Figueiredo ACMG, Gomes-Filho IS, Silva RB, *et al*. Maternal anemia and low birth weight: a systematic review and meta-analysis. *Nutrients* 2018;10:601.
18 Dadi AF, Miller ER, Bisetegn TA, *et al*. Global burden of antenatal depression and its association with adverse birth outcomes: an umbrella review. *BMC Public Health* 2020;20:173.
19 Lima SAM, El Dib RP, Rodrigues MRK, *et al*. Is the risk of low birth weight or preterm labor greater when maternal stress is experienced during pregnancy? A systematic review and meta-analysis of cohort studies. *PLOS ONE* 2018;13:e0200594.
20 Fleischer NL, Merialdi M, van Donkelaar A, *et al*. Outdoor air pollution, preterm birth, and low birth weight: analysis of the world health organization global survey on maternal and perinatal health. *Environ Health Perspect* 2014;122:425–30.
21 Flower A, Shawe J, Stephenson J, *et al*. Pregnancy planning, smoking behaviour during pregnancy, and neonatal outcome: UK millennium cohort study. *BMC Pregnancy Childbirth* 2013;13:238.
22 Jaddoe VWV, Troe E-JWM, Hofman A, *et al*. Active and passive maternal smoking during pregnancy and the risks of low birthweight and preterm birth: the generation R study. *Paediatr Perinat Epidemiol* 2008;22:162–71.
23 Patra J, Bakker R, Irving H, *et al*. Dose-response relationship between alcohol consumption before and during pregnancy and the risks of low birthweight, preterm birth and small for gestational age (SGA)-A systematic review and meta-analyses. *BJOG* 2011;118:1411–21.
24 Dos Santos JF, de Melo Bastos Cavalcante C, Barbosa FT, *et al*. Maternal, fetal and neonatal consequences associated with the use of crack cocaine during the gestational period: a systematic review and meta-analysis. *Arch Gynecol Obstet* 2018;298:487–503.
25 Hill A, Pallitto C, McCleary-Sills J, *et al*. A systematic review and meta-analysis of intimate partner violence during pregnancy and selected birth outcomes. *Int J Gynaecol Obstet* 2016;133:269–76.
26 Donovan BM, Spracklen CN, Schweizer ML, *et al*. Intimate partner violence during pregnancy and the risk for adverse infant outcomes: a systematic review and meta-analysis. *BJOG* 2016;123:1289–99.
27 Lyons RA, Jones KH, John G, *et al*. The sail databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 2009;9:3.
28 HMICFRS. Constabulary © her majesty's inspectorate of, fire. police effectiveness 2015 (vulnerability) – dyfed-powys police. n.d. Available: https://www.justiceinspectorates.gov.uk/hmicfrs/publications/police-effectiveness-vulnerability-2015-dyfed-powys/
29 Gelman A, Hill J. *Data analysis using regression and multilevel/hierarchical models*. 2006.
30 Lewis RJ. An introduction to classification and regression tree (CART) analysis. In: *Annual meeting of the society for academic emergency medicine in San Francisco*. California, 2000.
31 Atkinson Beth. Rpart function | R documentation. Available: https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart [Accessed 14 Jan 2021].
32 ONS. Birth characteristics in england and wales - office for national statistics. 2019. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthcharacteristicsinenglandandwales/2017 [Accessed 14 Jul 2021].
33 Regan AK, Ball SJ, Warren JL, *et al*. A population-based matched-sibling analysis estimating the associations between first interpregnancy interval and birth outcomes. *Am J Epidemiol* 2019;188:9–16.
34 Class QA, Rickert ME, Oberg AS, *et al*. Within-family analysis of interpregnancy interval and adverse birth outcomes. *Obstet Gynecol* 2017;130:1304–11.
35 Hanley GE, Hutcheon JA, Kinniburgh BA, *et al*. Interpregnancy interval and adverse pregnancy outcomes: an analysis of successive pregnancies. *Obstet Gynecol* 2017;129:408–15.
36 Hutcheon JA, Nelson HD, Stidd R, *et al*. Short interpregnancy intervals and adverse maternal outcomes in high-resource settings: an updated systematic review. *Paediatr Perinat Epidemiol* 2019;33:48–59.
37 Chamberlain C, O'Mara-Eves A, Porter J, *et al*. Psychosocial interventions for supporting women to stop smoking in pregnancy. *Cochrane Database Syst Rev* 2017;2:CD001055.
38 Howard LM, Khalifeh H. Perinatal mental health: a review of progress and challenges. *World Psychiatry* 2020;19:313–27.
39 Faruk A, Cahyono ES. Prediction and classification of low birth weight data using machine learning techniques. *Indonesian J Sci Technol* 2018;3:18. 10.17509/ijost.v3i1.10799. Available: http://ejournal.upi.edu/index.php/ijost/issue/view/IJoST%3A%20Volume%203%2C%20Issue%201%2C%202018
40 Loreto P, Peixoto H, Abelha A, *et al*. Predicting low birth weight babies through data mining. In: Rocha Á, Adeli H, Reis LP, eds. *New knowledge in information systems and technologies*. Cham: Springer International Publishing, 2019: 568–77.
41 Ahmadi P, Alavimajd H, Khodakarim S, *et al*. Prediction of low birth weight using random forest: A comparison with logistic regression. *Arch Adv Biosci* 2017;8:36–43.
42 Ghavi A, Fadakar Sogheh K, Niknamy M, *et al*. Investigating the relationship between maternal lifestyle during pregnancy and low-birth-weight of term neonates. *Iran J Obstet Gynecol Infertil* 2012;15:14–24.
43 Xi C, Luo M, Wang T, *et al*. Association between maternal lifestyle factors and low birth weight in preterm and term births: a case-control study. *Reprod Health* 2020;17:93.

My input

As the lead researcher for the 'Vulnerability Profiling' project and WECC Phase 4, I developed the research plan, including conducting the literature review and finalising the covariates and their sources from routine administrative data. After preparing the core dataset, I meticulously cleaned the linked routine data to ensure it was ready for analysis. I constructed the MLR and DT models using the R software package and SQL on the IBM DB2 platform within the SAIL Databank environment. The findings were presented in a journal paper I authored.

Impact

- This article has been published in The BMJ Open (*impact factor 2.4, cite score 4.4, acceptance rate 36%)*.
- The paper has been cited in seven other published works as mentioned by google scholar.
- The findings have been utilised as the outcome of the pilot program 'Vulnerability Profiling' conducted by Public Health Wales', serving as the foundation for their future research and initiatives
- The findings of the paper have been disseminated as a Data Insight report across the UK.
- The paper received extensive international media coverage, including
    1. Media Xpress (https://medicalxpress.com/news/2023-02-factors-birth-weight.html), USA
    2. Newswise (https://www.newswise.com/articles/new-study-identifies-risk factors-associated-with-low-birthweights), USA
    3. ScienceDaily (https://www.sciencedaily.com/releases/2023/02/230215143326.htm), USA
    4. News-Medical.net (https://www.news-medical.net/news/20230215/Researchers-gain-deeper-understanding-of-risk factors-associated-with-low-birthweights.aspx), Australia.
    5. Archynewsy (https://www.archynewsy.com/low-birth-weight-these-are-the-causes/), USA
    6. Italia Salute.it (https://www.italiasalute.it/6886/I-fattori-di-rischio-per-basso-peso-alla-nascita.html), Italy
    7. SciTechDaily(https://scitechdaily.com/new-study-identifies-5-key-factors-that-can-reduce-the-risk-of-low-birth-weight/#google_vignette), USA
- I was interviewed by That's TV and the segment aired on February 17th 2023 on Channel 8.

Conclusion

The findings from this chapter highlight the profound impact of maternal health and socio-economic factors on birth outcomes. LBW is not only a health concern at birth but has long-term consequences for childhood development and educational attainment. Understanding the broader implications of LBW is crucial for informing policies aimed at reducing health disparities and improving child outcomes. By utilising routine administrative data, this study provides robust evidence to support targeted interventions that enhance maternal health services and prenatal care programs. I have established the relevance of my work and its contribution to population health and social care research. The impact of early-life health factors extends beyond infancy and childhood into academic performance and long-term educational success. The next chapter delves into school readiness, examining how birth-related vulnerabilities, socio-economic conditions and health factors contribute to disparities in educational attainment.

Chapter 3: Factors associated with low school readiness, a linked health and education data study in Wales, UK

Critical summary

Background

As children grow older, they prepare to enter school, where school readiness encompasses cognitive, social and emotional preparedness. Research has established the detrimental effects of low school readiness, which can significantly increase a child's vulnerability and impact future outcomes (5,57). Identifying the risk factors associated with low school readiness is therefore a priority for public health systems and early-life policy approaches in the UK, as it helps reduce the burden on public health. The paper, '*Factors associated with low school readiness, a linked health and education data study in Wales, UK*' published in *PLOS One*, aims to develop a comprehensive understanding of the risk factors contributing to low school readiness by utilising linked administrative data.

Utilisation of administrative data

The existing literature on identifying risk factors associated with low school readiness has primarily concentrated on survey-based data collection methods (58–60). However, acquiring survey data is often costly and resource-intensive, requiring significant financial and logistical investments (61). Additionally, such methods face challenges including sampling bias (which may fail to represent the general population) (62), recall bias (due to inaccurate self-reported information) (63), limited depth of data and temporal constraints (64).

In response to these challenges, this study adopts a novel approach to measuring school readiness using routine data. This study, the first of its kind, has built a linked routine data framework for a nationally representative population sample. By utilising data collected through standard educational assessments and administrative records, this method not only reduces costs associated with conducting a survey-based research work but also minimises the risk of selection and recall biases. Furthermore, it enhances the feasibility of large-scale research and provides a more accessible means of identifying risk factors to improve school readiness outcomes.

- *Outcome variable:* In this study, school readiness was assessed using the Pre16 Education Attainment dataset, which comprises routine administrative data. A binary Foundation Phase Indicator was created to measure school readiness for children aged 6 or 7. This indicator reflects whether a child has achieved at least

the expected level 5 or higher in early learning goals across three areas: i) personal and social development, wellbeing and cultural diversity; ii) language, literacy and communication skills in English/Welsh; and iii) mathematical development.

- *Risk factors:* All risk factors identified through the literature review were derived from routinely collected administrative data. These include socio-economic deprivation (from WDS), the child's birth information (from NCCHD), both child and maternal physical and mental health data (from EDDS, PEDW and WLGP). Additionally, maternal lifestyle factors were sourced from PEDW, WLGP and SMD.

The use of anonymised data linkage through the ALF facilitated the integration of these datasets, effectively connecting each child with their maternal and family records. This comprehensive approach enabled the longitudinal tracking of the study population from birth through their foundation phase, enhancing both the robustness and relevance of the findings. By leveraging extensive, linked administrative data, this study addresses critical gaps in understanding school readiness. The findings provide valuable insights into the predictors of low school readiness, offering evidence that can inform future policy and practice to better support children and families.

Application of data science methods

The introduction of this thesis outlines a methodological framework that integrates data science with traditional statistical models and data-driven machine learning approaches. The primary objective of this research is to identify the most significant risk factors associated with low school readiness. While existing studies have developed models to investigate the predictors of low school readiness, they have often been constrained by the absence of granular level maternal and child health records (65,66). The lack of these critical data points can compromise the accuracy and reliability of the findings, limiting the ability to draw meaningful conclusions about the factors influencing school readiness. To address this limitation, the present study employs multivariable LR models and DT models using a robust routine dataset with more granular level exposure data. This approach aims to provide a more nuanced understanding of the predictors of low school readiness. The use of MLR enables the simultaneous examination of multiple risk factors, providing insights into their relative contributions to school readiness outcomes. Meanwhile, DT models provide a visual representation of the decision-making process, highlighting the most significant predictors and their interactions. By combining these approaches, this research enhances the validity of its findings and facilitates the identification of targeted interventions to effectively address the identified risk factors.

Early-life vulnerability profiling

This research paper aims to make significant contributions to the field of early-life vulnerability profiling by developing a linked data approach that integrates numerous administrative datasets. This methodology enables a more comprehensive

understanding of the factors contributing to low school readiness, which is essential for developing effective interventions.

This study provides a holistic understanding of the vulnerability profile of the children who are at high risk of low school readiness. A key contribution of the study is the identification of modifiable risk factors associated with low school readiness. The findings emphasise the critical roles of demographic factors, socio-economic status, maternal health and early-life experiences in determining a child's readiness for school. It suggests that deprivation is one of the main risk factors for low school readiness even after adjusting for maternal and child's health. Results showed that boys and those with poor school attendance, are at greater risk of low school readiness. Additionally, the study highlights the importance of supporting families dealing with learning difficulties and illnesses, such as epilepsy, to further help with a child's school readiness. Overall, these efforts could strengthen lifelong learning foundations and reduce educational inequalities.

This work provides a novel holistic perspective on school readiness and the results align with existing literature. For instance, in the thesis published in 2024, Amanda Sanz demonstrated that the socio-economic background of children significantly influences their academic and cognitive development (67). Similarly, quantitative and qualitative research works conducted by Belkacem et al. (2024), Mensah et al. (2010) and Puha et al. (2016) highlight the substantial impact of maternal physical and mental health during and after pregnancy on a child's cognitive developmental outcomes (68–70). Mensah et al. used MCSD data to establish maternal physical health as a positive predictor of child's learning development, even after adjusting for maternal psychological distress (69). A systematic review conducted by Phua et al. suggested that positive maternal mental health contributes to children's overall development (70). While existing literature often investigates risk factors in isolation, for example, considering maternal health as a predictor of child development, maternal lifestyle factors (substance use and alcohol consumption), or area-level deprivation, a systematic and holistic understanding of the combination of these risk factors were mostly unavailable in previous studies. This work addresses these gaps and adds to the novel contribution of the research. The present study employs a robust statistical framework, utilising MLR to investigate the associations between risk factors and school readiness outcomes. This approach enables the simultaneous examination of multiple risk factors, providing insights into their relative contributions. Additionally, DT models further enhance the analysis by visually representing the hierarchical risk factor clusters and identifying the most common and significant predictors. The findings have profound implications for policymakers and practitioners in early-life education and public health. The study underscores the necessity of targeted interventions that address the identified risk factors, particularly among vulnerable populations.

Published journal paper

PLOS ONE

RESEARCH ARTICLE

Factors associated with low school readiness, a linked health and education data study in Wales, UK

Amrita Bandyopadhyay[1]*, **Emily Marchant**[1], **Hope Jones**[1], **Michael Parker**[1], **Julie Evans**[2], **Sinead Brophy**[1]

1 National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Wales, United Kingdom, **2** Public Health Wales, Keir Hardie University Health Park, Wales, United Kingdom

*[redacted]

Citation: Bandyopadhyay A, Marchant E, Jones H, Parker M, Evans J, Brophy S (2023) Factors associated with low school readiness, a linked health and education data study in Wales, UK. PLoS ONE 18(12): e0273596. https://doi.org/10.1371/journal.pone.0273596

Editor: Sreeram V. Ramagopalan, University of Oxford, UNITED KINGDOM

Received: August 12, 2022

Accepted: November 20, 2023

Published: December 11, 2023

Data Availability Statement: The data is held in the Secure Anonymised Information Linkage Databank (Data Science Building, Swansea University, Singleton Park, SA28PP) TRE (Trusted Research Environments) and is available through application. The data is restricted and requires review by Information Governance Review Panel (IGRP) who are providing independent guidance and advice on Information Governance policies, procedures and processes for SAIL Databank. All necessary information will be available on the following link: https://saildatabank.com/contact/.

Abstract

Background

School readiness is a measure of a child's cognitive, social, and emotional readiness to begin formal schooling. Children with low school readiness need additional support from schools for learning, developing required social and academic skills, and catching-up with their school-ready peers. This study aims to identify the most significant risk factors associated with low school readiness using linked routine data for children in Wales.

Method

This was a longitudinal cohort study using linked data. The cohort comprises of children who completed the Foundation Phase assessment between 2012 and 2018. Individuals were identified by linking Welsh Demographic Service and Pre16 Education Attainment datasets. School readiness was assessed via the binary outcome of the Foundation Phase assessment (achieved/not achieved). This study used multivariable logistic regression model and a decision tree to identify and weight the most important risk factors associated with low school readiness.

Results

In order of importance, logistic regression identified maternal learning difficulties (adjusted odds ratio 5.35(95% confidence interval 3.97–7.22)), childhood epilepsy (2.95(2.39–3.66)), very low birth weight (2.24(1.86–2.70), being a boy (2.11(2.04–2.19)), being on free school meals (1.85(1.78–1.93)), living in the most deprived areas (1.67(1.57–1.77)), maternal death (1.47(1.09–1.98)), and maternal diabetes (1.46(1.23–1.78)) as factors associated with low school readiness. Using a decision tree, eligibility for free school meals, being a boy, absence/low attendance at school, being born late in the academic year, being a low birthweight child, and not being breastfed were factors which were associated with low school readiness.

Funding: This work was funded by Public Health Wales (PHW), grant number (105186). This research has been carried out as part of the ADR Wales programme of work. The ADR Wales programme of work is aligned to the priority themes as identified in the Welsh Government's national strategy: Prosperity for All. ADR Wales brings together data science experts at Swansea University Medical School, staff from the Wales Institute of Social and Economic Research, Data and Methods (WISERD) at Cardiff University and specialist teams within the Welsh Government to develop new evidence which supports Prosperity for All by using the SAIL Databank at Swansea University, to link and analyse anonymised data. ADR Wales is part of the Economic and Social Research Council (part of UK Research and Innovation) funded ADR UK (grant ES/S007393/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Conclusion

This work suggests that public health interventions focusing on children who are: boys, living in deprived areas, have poor early years attendance, have parents with learning difficulties, have parents with an illness or have illnesses themselves, would make the most difference to school readiness in the population.

Introduction

Background

Early childhood education shapes the direction of a child's development, enhances their ability to learn in the school environment and strengthens their foundation for lifelong learning [1,2]. School readiness encompasses cognitive, social, and emotional aspects and indicates if a child can achieve at an appropriate level in formal school. School readiness is also a determinant of health and wellbeing over the life course [3,4]. It is strongly linked to the pre-school environment, and it indicates the acquisition of the necessary social skills, emotional skills, knowledge, and attitude to effectively engage and learn in school. School readiness is defined by a child's physical well-being and motor development (e.g., co-ordination, fine motor-skills), social and emotional development (co-operation, empathy, and the ability to express their emotion), approaches towards learning (enthusiasm, curiosity, temperament), language and communication (listening and speaking), basic knowledge (essential vocabulary and numbers) and cognitive skills (problem solving) [4].

A review of published literature on the risk factors associated with school readiness indicates that area-level characteristics, parental demography, and parental and child health conditions play a significant role in school readiness. Factors associated with higher school readiness include higher levels of child care provision in the area where the child is brought up [5,6], living in private housing [7], the mother's age (between late twenties or thirties) [7,8], breastfeeding (higher rates and longer duration) [7,9], dual parent households, a nurturing parenting style [7,10] and parents with good physical [10,11] and mental health [7,9,10]. Similarly, good physical health of the child (being born at term and a healthy birth weight) [12,13] is also associated with higher school readiness. Conversely, low access to childcare, higher levels of unemployment (area and family level), living in social housing, exposure to poor environment such as damp, maternal heavy drinking behaviours [14], mother who smoked during pregnancy [5,12], teenage mothers or older mothers (35+ years) and parents with poor physical health (hypertension, diabetes), poor mental health, single parent or step-parent families, low expectations by the parent for the child, preterm or low birth weight child, and poor health of the child are also associated with low school readiness [6,7]. An Australian data linkage study conducted by Chittleborough et al, identified a group of predictors (such as maternal age, smoking during pregnancy, parity, marital status, and both parents' occupation and gender) which were capable to identify the children at risk of developmental vulnerability at school entry [15].

Since being school ready is associated with many positive outcomes, improving school readiness is a necessary strategy for economic development and social mobility [16]. If children are not school ready, it can take many years for them to catch up with their peers, if ever, [17,18] and therefore contribute to widening inequalities. School readiness has been identified as a key public health concern in a recent review of UK public health systems and policy approaches to early child development [19]. It is very challenging to identify the right individuals (children and families) who are at risk in order to provide the necessary support [20].

Therefore, identifying the most significant risk factors is a priority in closing the gap in children's school readiness and improving outcomes for children. Studies have shown that routine data obtained during a child's birth can help to identify the children and the families at risk of poor development [21,22]. A framework using routinely collected administrative data can inform the appropriate supporting agencies to provide adequate help and support to the most vulnerable of the society.

Objective

The aim of the study is to identify and weight the most significant risk factors of low school readiness using linked routine data for children in Wales. This work also examined the risk factors which were clustered together and build a vulnerability profile of the children who are at risk of low school readiness. The factors which are associated with school readiness are examined using; a) traditional statistical methods (multivariable logistic regression model, to observe the highly associated risk factors) and b) data driven supervised machine learning classification algorithm (decision tree, to measure the commonly observed and prevalent risk factors at the population level). In a logistic regression model, the log-odds for low school readiness as a linear combination of explanatory variables and confounders have been investigated. On the other hand, the decision tree model, based on recursive partitioning, highlights the statistically significant hierarchically clustered features for low school readiness and captures complex relationship between the risk factors and low school readiness. The hierarchically clustered features from the decision tree and the risk factors identified from the logistic model are important to cross-validate the set of overlapping risk factors and serves to strengthen the importance of the findings. These risk factors will inform the development of a profiling model by identifying the socio-economic and physical and mental health barriers that the child and/or their families face, which may impact the child's ability to meet the developmental milestones necessary to progress effectively through the early years. The primary focus of the model is to build a holistic understanding of the most significant risk factors of the low school readiness which can inform the necessary support system required for the individuals and families at highest need and make more efficient use of the resources.

Methods

Sample selection and data linkage

In this cohort study, the study population was derived by linking the Welsh Demographic Service (WDS) dataset (administrative dataset about individuals in Wales that use NHS services) and the Pre16 Education Attainment dataset (individual-level administrative data relating to the education system in Wales). The study population consists of children who completed Foundation Phase (a statutory curriculum for children aged 3–7 years) [23] between 2012 and 2018. The data linkage was done using an encrypted key known as Anonymised Linking Fields (ALF) in the Secure Anonymised Information Linkage (SAIL) Databank [24,25]. Residential Anonymised Linking Fields (RALFs) are an encrypted residential address available in WDS dataset, which is also linked with a smaller geographical unit known as lower super output area (LSOA). Using ALFs, RALFs and LSOA the study population were anonymously linked with the individuals living with the child in the same household during the child's Foundation Phase [26]. Children without valid and continuous RALFs and primary care records in Welsh Longitudinal General Practice (WLGP) dataset in SAIL until their completion of Foundation Phase were not included in the study to ensure the complete coverage of exposure and outcome data during the study period. The study population was linked with the National Community Child Health Database (NCCHD) to obtain birth and maternal records during

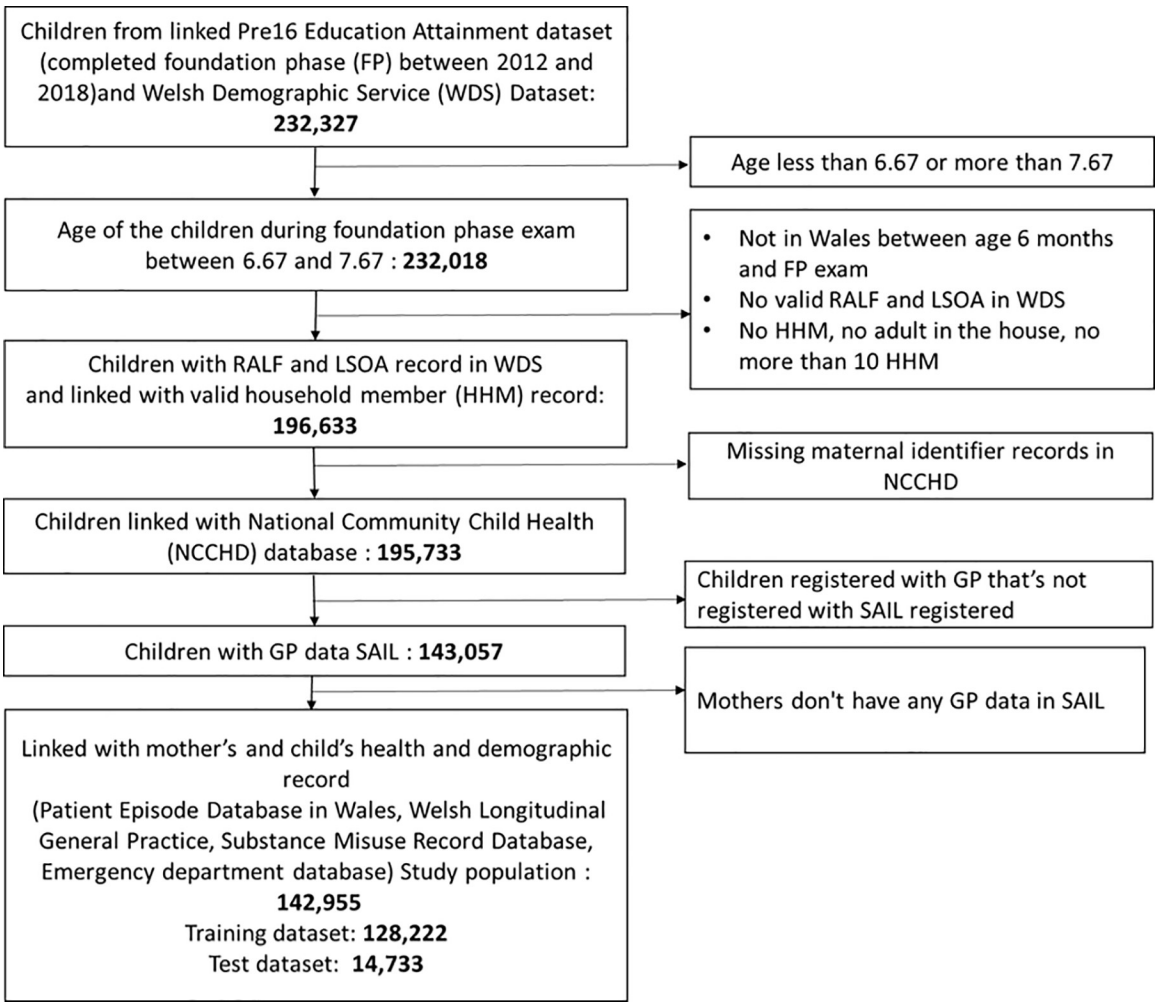


Fig 1. Flow diagram of the study population.

https://doi.org/10.1371/journal.pone.0273596.g001

childbirth. Records with missing maternal identifiers and mothers with no primary care record in WLGP dataset were not included in the study. The flow diagram of the selection of the study population is presented in Fig 1.

Risk factors from routine data

The selection of risk factors associated with low school readiness has been informed by the literature review undertaken at the inception of the study. The risk factors had been selected from the routinely collected electronic administrative and health datasets and this provided the framework upon which the current study was developed. The literature review focused on observational studies including case controls, cohort studies and studies using linked routine data with the primary or secondary outcomes examining school readiness. Depending on the strength of association (Odds Ratio) between the risk factors and low school readiness a list of risk factors were prepared, and their analogous variables were created or selected from linked routine data. The literature review to select the risk factors and how these were mapped with routine data, have been described in a Supplementary document (Appendix 1 in S1 File and Appendix 2 in S1 File). General demography and birth-related variables including gender,

gestational age, birth weight, breastfeeding, mode of delivery (caesarean section/assisted delivery/natural delivery) and maternal age at childbirth were obtained from NCCHD. The multiple birth (singleton/non-singleton) flag was derived using week of birth of the child, encrypted maternal identifier and the birth order of the children. To identify the children who lost their mother before the Foundation Phase, a binary variable was derived. Maternal physical health (diabetes, cancer, anaemia, hypertension, learning difficulty) and mental health (depression, anxiety, serious mental illness, medication related to anxiety/depression) related primary and secondary care records during and after pregnancy until Foundation Phase were obtained from WLGP and hospital admission dataset—Patient Episode database in Wales (PEDW). The Substance Misuse Database (SMD) was used to populate information on the mothers' alcohol or other substance abuse related record during the study period. Any coded READ and ICD10 codes related to substance misuse on WLGP and PEDW dataset were also considered in this study. Mothers' alcohol related hospital admission records were obtained from PEDW. Maternal smoking during and after pregnancy were obtained from WLGP, smoking related READ code mentioned on the dataset during the study period were considered to build the variable. The record of physical assault related hospital admissions of mothers during or post pregnancy was obtained from PEDW. The hospital admission and GP records of the children for epilepsy, asthma, diabetes, ear infections, and eye infections were considered as a measure of child health conditions. Any emergency hospital admission and any accident and emergency (A&E) attendance of the study population between birth and Foundation Phase were obtained from PEDW and Emergency Department dataset (EDDS). READ code version 2 and ICD10 codes have been used to identify the health records from WLGP and PEDW dataset (see Appendix 3 in S1 File). Any coded diagnosis of the above-mentioned physical and mental health conditions for mother and child during their GP visit (obtained from WLGP) or hospital admission (obtained from PEDW) were considered in this study. The children's age at the completion of their Foundation Phase and the total number of days they were absent in the school in early years (e.g., nursery) were obtained from Pre16 Education Attainment dataset. Household characteristics such as living in a single adult household, total number of adults, and total number of other children in the household were derived from the WDS dataset. In this study the eligibility for free school meals (FSM) during Foundation Phase was used to measure the family-level deprivation of the study population. The area-level deprivation was measured by the Welsh Index of Multiple Deprivation (WIMD) 2014 which provides a measure of the relative deprivation in Wales linked to LSOA [27]. The local authorities and the type of local area (urban/rural) where the children were brought up during Foundation Phase were included in the study.

School readiness from routine data

The binary Foundation Phase Indicator variable was obtained from the Pre16 Education Attainment dataset and was used as a measure of school readiness from the routine data in the current study. The National Curriculum assess school readiness using the Foundation Phase Indicator at the end of early year foundation stage where the child would be at the age of 6 or 7. The Foundation Phase Indicator represents whether the child has achieved at least the expected level 5 or above in the early stage learning goals in the following areas;—i) personal and social development, well-being and cultural diversity, ii) language, literacy, and communication skills–English/Welsh and iii) mathematical development [28]. In this study a binary variable has been derived based on the Foundation Phase Indicator record as a measure of school readiness from routine data.

- Low school readiness = Not achieved Foundation Phase

- School readiness = Achieved Foundation Phase

Statistical analysis

A multivariable logistic regression model was first developed to identify and weight the most important risk factors associated with school readiness. Next, we built a data driven machined learning classifier model using decision tree to investigate the most commonly observed risk factors at the population level. Since the children with learning difficulties or special educational needs tend to have a much higher risk of low school readiness, they were removed from the models. Data preparation including data linkage was performed on DB2 SQL platform and the statistical analysis was done in R version 4.0.3.

Logistic regression. To identify the most important risk factors associated with low school readiness we used multivariable logistic regression. Variables included gender, gestational age, birthweight, breastfeeding, caesarean section, multiple birth, maternal age, maternal death before Foundation Phase, maternal physical and mental health, child physical (epilepsy, asthma, diabetes, ear, and eye) and mental health conditions (depression, anxiety), free school meal uptake, local area status and number of adults and children living in the same household. The significant risk factors of low school readiness are presented with their adjusted Odds Ratio (aOR) and 95% confidence interval (CI).

Decision tree. A classification tree–decision tree algorithms were developed using RPART (Recursive Partitioning And Regression Trees) packages in R [29,30]. The algorithm repeatedly partitions the data into multiple sub-spaces to reach the homogeneous end subspace, hence it is called recursive partitioning. For decision trees, the data for one representative local authority was removed from the dataset and used as the testing dataset to validate the model performance and examine generalisability within areas of Wales.

Results

Overall sample characteristics

The study population consisted of 142,955 children (training dataset: 128,222, testing dataset: 14,733) who completed Foundation Phase between 2012 and 2018 in Wales (see Table 1). 14.32% (Training dataset: 14.15%, Testing dataset: 15.75%) children did not achieve in Foundation Phase. The study population consisted of 51.24% boys, 42.87% were not breastfed and 24.83% were born via caesarean section. 8.33% were born to mothers aged below 19, 0.14% of mothers had learning difficulties and 0.23% lost their mother before their Foundation Phase assessment. There were 0.1% mothers who had an alcohol related hospital admission, 0.36% with substance abuse and 14.63% had a smoking record in WLGP during pregnancy. 0.64% of children had an epilepsy related GP visit, 0.46% had a hospital admission record for epilepsy. 3.30% and 4.54% children were admitted to hospital for asthma and ear infection respectively before they completed Foundation Phase. 56.37% of children had at least one emergency hospital admission and 66.4% had A&E records anytime between birth and Foundation Phase. 0.90% of children (Training dataset: 0.88%, Testing dataset 1.07%) were diagnosed with a learning difficulty. 22.05% of children were in single adult households, 19.57% were eligible for FSM and 25.66% lived in most deprived area measured by WIMD. Overall characteristics of the study population have been described in Table 1.

Logistic regression results. Significant risk factors associated with low school readiness included: maternal learning difficulty (aOR (95% CI): 5.35 (3.97–7.22)), child epilepsy (2.95 (2.39–3.66)), having a very low birthweight (2.24 (1.86–2.70)), boys (2.11 (2.04–2.19)), being eligible for FSM (1.85 (1.78–1.93)), being extremely preterm (1.41 (1.04–1.91)), living in the

Table 1. Characteristics of the study population.

Variables	Overall (n = 142,955)		Training dataset (n = 128,222)		Testing dataset (n = 14,733)	
Gender						
Girl	69,703	48.76%	62,420	48.68%	7,283	49.43%
Boy	73,252	51.24%	65,802	51.32%	7,450	50.57%
Gestational age						
Extremely pre-term: <28 weeks	358	0.25%	319	0.25%	39	0.26%
Very pre-term: 28–31	1,173	0.82%	1,017	0.79%	156	1.06%
Pre-term: 32–36	8,434	5.90%	7,524	5.87%	910	6.18%
Term: 37–42	131,249	91.81%	117,708	91.80%	13,541	91.91%
Late term: 43–45	899	0.63%	849	0.66%	50	0.34%
Unknown/NULL	842	0.59%	805	0.63%	37	0.25%
Birth weight						
Very low: <1500 g	1,454	1.02%	1,295	1.0%	159	1.08%
Low: 1500-<2500	8,185	5.73%	7,207	5.6%	978	6.64%
Normal: 2500-<4000g	115,844	81.04%	103,767	80.9%	12,077	81.97%
High: 4000-5000g	16,802	11.75%	15,320	11.9%	1,482	10.06%
Unknown	670	0.47%	633	0.5%	37	0.25%
Breastfeeding						
No	61,287	42.87%	54,838	42.77%	6,449	43.77%
Yes	73,988	51.76%	66,037	51.50%	7,951	53.97%
Unknown	7,680	5.37%	7,347	5.73%	333	2.26%
C-section birth						
	35,489	24.83%	31,275	24.39%	4,214	28.60%
Multiple birth						
Non-singleton	3,922	2.74%	3,546	2.77%	376	2.55%
Maternal age						
Less than 19	11,910	8.33%	10,416	8.12%	1,494	10.14%
20–24	32,384	22.65%	28,598	22.30%	3,786	25.70%
25–29	39,356	27.53%	35,093	27.37%	4,263	28.94%
30–34	35,840	25.07%	32,534	25.37%	3,306	22.44%
35 and above	23,458	16.41%	21,574	16.83%	1,884	12.79%
Unknown	7	0.00%	7	0.01%		
Death of mother before Foundation Phase						
	327	0.23%	286	0.22%	41	0.28%
Diabetes PEDW (mother)						
	1,462	1.02%	1,308	1.02%	154	1.05%
Diabetes GP (mother)						
	1,367	0.96%	1231	0.96%	136	0.92%
Cancer PEDW (mother)						
	1,192	0.83%	1,097	0.86%	95	0.64%
Cancer GP (mother)						
	1,037	0.73%	947	0.74%	90	0.61%
Anaemia PEDW (mother)						
	7,317	5.12%	6,799	5.30%	518	3.52%
Anaemia GP (mother)						
	15,680	10.97%	14,340	11.18%	1,340	9.10%
Hypertension GP (mother)						

(*Continued*)

Table 1. (Continued)

Variables	Overall (n = 142,955)		Training dataset (n = 128,222)		Testing dataset (n = 14,733)	
	2,599	1.82%	2,330	1.82%	269	1.83%
Learning Difficulty GP (mother)						
	205	0.14%	182	0.14%	23	0.16%
Depression PEDW (mother)						
	5,179	3.62%	4,648	3.62%	531	3.60%
Depression GP (mother)						
	29,332	20.52%	26,060	20.32%	3,272	22.21%
Anxiety PEDW (mother)						
	2,913	2.04%	2,626	2.05%	287	1.95%
Anxiety GP (mother)						
	30,278	21.18%	26,916	20.99%	3,362	22.82%
Anti-Depression/anxiety medication (mother)						
	396	0.28%	359	0.28%	37	0.25%
Serious Mental Illness PEDW (mother)						
	710	0.50%	624	0.49%	86	0.58%
Serious Mental Illness GP (mother)						
	776	0.54%	676	0.53%	100	0.68%
Alcohol PEDW (mother)						
During pregnancy	137	0.10%	129	0.10%	8	0.05%
After pregnancy	1,362	0.95%	1,232	0.96%	130	0.88%
Smoking GP (mother)						
During pregnancy	20,913	14.63%	18,720	14.60%	2,193	14.88%
After pregnancy	37,142	25.98%	33,477	26.11%	3,665	24.88%
Substance misuse (any) SMD (mother)						
During pregnancy	167	0.12%	149	0.12%	18	0.12%
After pregnancy	2,084	1.46%	1,823	1.42%	261	1.77%
Substance misuse (other drug) PEDW (mother)						
During pregnancy	272	0.19%	249	0.19%	23	0.16%
After pregnancy	1,355	0.95%	1,233	0.96%	122	0.83%
Substance misuse (other drug) GP (mother)						
During pregnancy	511	0.36%	459	0.36%	52	0.35%
After pregnancy	1,917	1.34%	1,705	1.33%	212	1.44%
Assault PEDW (mother)						
	572	0.40%	514	0.40%	58	0.39%
Diabetes PEDW (child)						
	212	0.15%	184	0.14%	28	0.19%
Diabetes GP (child)						
	199	0.14%	178	0.14%	21	0.14%
Epilepsy PEDW (child)						
	652	0.46%	554	0.43%	98	0.67%
Epilepsy GP (child)						
	916	0.64%	816	0.64%	100	0.68%
Asthma PEDW (child)						
	4,719	3.30%	4249	3.31%	470	3.19%
Asthma GP (child)						
	55,001	38.47%	49,468	38.58%	5,533	37.56%

(*Continued*)

Table 1. (Continued)

Variables	Overall (n = 142,955)		Training dataset (n = 128,222)		Testing dataset (n = 14,733)	
Ear PEDW (child)						
	6,493	4.54%	5870	4.58%	623	4.23%
Eye PEDW (child)						
	3,836	2.68%	3432	2.68%	404	2.74%
Any emergency hospital admission (child)						
	80,588	56.37%	71282	55.59%	9306	63.16%
Any A&E attendance (child)						
	94,924	66.40%	84818	66.15%	10106	68.59%
Learning Difficulty (child)						
	1,290	0.90%	1132	0.88%	158	1.07%
Low School readiness						
Did not achieve Foundation Phase	20,468	14.32%	18,148	14.15%	2,320	15.75%
Free school meal						
	27,971	19.57%	24,587	19.18%	3,384	22.97%
WIMD 2014—overall						
1 (most deprived)	36,682	25.66%	32,237	25.14%	4,445	30.17%
2	30,647	21.44%	25,862	20.17%	4,785	32.48%
3	26,486	18.53%	24,531	19.13%	1,955	13.27%
4	22,283	15.59%	21,020	16.39%	1,263	8.57%
5 (least deprived)	26,857	18.79%	24,572	19.16%	2,285	15.51%
Local area—urban/rural						
Rural town	22,578	15.79%	18,592	14.50%	3,986	27.05%
Rural village	13,494	9.44%	13,418	10.46%	76	0.52%
Urban city and town	106,883	74.77%	96,212	75.04%	10,671	72.43%
No of adult in the household						
1	31,524	22.05%	27809	21.69%	3,715	25.22%
2	83,698	58.55%	75271	58.70%	8,427	57.20%
3	17,360	12.14%	15656	12.21%	1,704	11.57%
4 or above	10,373	7.26%	9486	7.40%	887	6.02%
No of children in the household (excluding the cohort member)						
0	23,706	16.58%	21,005	16.38%	2,701	18.33%
1	67,693	47.35%	60,491	47.18%	7,202	48.88%
2	33,989	23.78%	30,720	23.96%	3,269	22.19%
3	11,714	8.19%	10,614	8.28%	1,100	7.47%
4 or above	5,853	4.09%	5,392	4.21%	461	3.13%

Descriptive statistics of the study population stratified by their school readiness has been included as a supplementary file (please see Appendix 4 in S1 File).

https://doi.org/10.1371/journal.pone.0273596.t001

most deprived area (1.67 (1.57–1.77)), not being breastfed (1.25 (1.21–1.30)), maternal death (1.47 (1.09–1.98)), maternal diabetes (1.46 (1.23–1.78)), smoking in pregnancy (1.36 (1.30–1.43)), child hospital admissions/illness for asthma (1.12 (1.03–1.22)), ear (1.36 (1.26–1.45)) and eye problems (1.30 (1.18–1.42)), single adult household (1.08 (1.04–1.12)), living with more than 3 children (1.63 (1.52–1.75)) in the household. The risk factors with their OR and upper and lower CI are presented in Table 2.

Result from decision tree

The training model consisted of 127,090 individuals who lived in Wales (excluding testing dataset). The most important variables in the model were: FSM, gender (boy), number of school absences, child's age while completing Foundation Phase, children with any emergency hospital admission, children with any A&E attendance, children with asthma, low birth weight, maternal substance misuse related GP record, maternal substance misuse related hospital admission, not being breastfed, children with ear problems and number of children in the household (higher number). The final decision tree model has been shown in Fig 2. Here are some case studies of the branches described in the decision tree model.

1. IF children are eligible for FSM (higher family level deprivation) -> Gender- Boys -> Total number of absent sessions more than 102 THEN the probability of Failed is 73% (terminal node 31)
2. IF children are not eligible for FSM (lower family level deprivation) -> Gender- Boys -> Younger in academic year ->Total number of absent sessions more than 82 THEN they are more likely to be Failed (terminal node 95).
3. IF children are eligible for FSM (higher family level deprivation) -> Gender- Girls -> Total number of absent sessions more than 84 THEN they are more likely to be Failed (terminal node 55).
4. IF children are eligible for FSM (higher family level deprivation) -> Gender- Boys -> Younger in academic year -> Low birth weight baby -> Not breastfed THEN they are more likely to be Failed (terminal node 119).
5. IF children are not eligible for FSM (lower family level deprivation) -> Gender- Girls THEN they are more likely to be Achieved (terminal node 4).
6. IF children are not eligible for FSM (lower family level deprivation) -> Gender- Girls -> Total number of absent sessions more than 41 THEN they are more likely to be Achieved (terminal node 10).

There were 14,575 children in the testing dataset. The model performance has been explained with the help of a confusion matrix. The model achieves 85.21% accuracy, 4.94% sensitivity, 99.37% specificity, 58.06% positive predictive values and 85.56% negative predictive value and 15% prevalence (see Tables 3 and 4).

Discussion

This study investigated the risk factors associated with low school readiness and developed two holistic models on a national level routine data framework. Here the multivariable regression model helped to identify the risk factors with the highest association/Odds Ratio but might not be common or frequently observed on a population level, the decision tree on the other hand contributed to identify the most important and common/frequent risk factors. Infrequent but highly associated events/factors which affect a child's school readiness include if the mother has a learning disability (0.14%), the child has epilepsy (0.64%) or is born extremely low birth weight (1%). However, there were also factors which were both highly associated and common such as being a boy (51%), where the odds of not being school ready is 2.11 than a girl (aOR more than twice that of girls), family level deprivation (eligible for FSM) which includes 19.5% of children, doubles the risk that they will not be school ready (aOR: 1.85). Low school attendance in early years (e.g., nursery) is associated with being 2% less likely to be school ready for every day missed in nursery.

Table 2. Logistic regression model to identify the risk factors associated with low school readiness.

Variable name	OR	Lower CI	Upper CI	P value
Gender				
Boy	2.11	2.04	2.19	0.00000
Gestational age (between 22 and 45)				
Extremely pre-term: <28 weeks	1.41	1.04	1.91	0.02527
Very pre-term: 28–31	0.98	0.81	1.19	0.87621
Pre-term: 32–36	1.03	0.96	1.11	0.40224
Late term: 43–45	1.18	0.98	1.43	0.08157
Unknown/NULL	0.95	0.75	1.20	0.65316
Birth weight (BW) (max 5000)				
Very low: <1500 g	2.24	1.86	2.70	0.00000
Low: 1500–<2500g	1.55	1.44	1.67	0.00000
High: 4000–5000g	0.85	0.80	0.90	0.00000
Unknown	1.01	0.77	1.33	0.93332
Breastfeeding				
No	1.25	1.21	1.30	0.00000
Unknown	1.29	1.19	1.39	0.00000
C-section birth				
	1.00	0.96	1.04	0.97489
Multiple birth				
Non-singleton	0.94	0.84	1.04	0.20643
Maternal age (between 10 and 65)				
Less than 19	1.22	1.14	1.30	0.00000
20–24	1.14	1.09	1.20	0.00000
25–29	1.05	1.00	1.10	0.04689
35 and above	1.11	1.05	1.17	0.00035
Unknown	1.64	0.26	10.14	0.59509
Death of mother				
	1.47	1.09	1.98	0.01154
Diabetes PEDW (mother)				
	1.00	0.84	1.19	0.99550
Diabetes GP (mother)				
	1.46	1.23	1.74	0.00002
Cancer PEDW (mother)				
	0.99	0.74	1.33	0.95178
Cancer GP (mother)				
	0.81	0.59	1.12	0.20313
Anaemia PEDW (mother)				
	0.94	0.87	1.02	0.12312
Anaemia GP (mother)				
	1.00	0.95	1.05	0.87589
Hypertension GP (mother)				
	1.02	0.91	1.16	0.69703
Learning Difficulty GP (mother)				
	5.35	3.97	7.22	0.00000
Depression PEDW (mother)				
	1.06	0.98	1.15	0.13704
Depression GP (mother)				

(*Continued*)

Table 2. (Continued)

Variable name	OR	Lower CI	Upper CI	P value
	1.13	1.09	1.18	0.00000
Anxiety PEDW (mother)				
	1.06	0.95	1.18	0.28780
Anxiety GP (mother)				
	0.98	0.94	1.02	0.30464
Anti Dep medication (mother)				
	0.89	0.67	1.18	0.41647
Serious Mental Illness PEDW (mother)				
	1.00	0.81	1.24	0.97260
Serious Mental Illness GP (mother)				
	1.00	0.82	1.23	0.97759
Alcohol PEDW (mother)				
During pregnancy	1.44	0.97	2.13	0.07261
After pregnancy	0.98	0.84	1.13	0.74478
Smoking GP (mother)				
During pregnancy	1.36	1.30	1.43	0.00000
After pregnancy	1.29	1.24	1.34	0.00000
Substance misuse (any) SMD (mother)				
During pregnancy	1.00	0.67	1.49	0.99571
After pregnancy	1.17	1.03	1.32	0.01335
Substance misuse (other drug) PEDW (mother)				
During pregnancy	1.35	1.00	1.81	0.04916
After pregnancy	1.05	0.90	1.22	0.52958
Substance misuse (other drug) GP (mother)				
During pregnancy	1.31	1.04	1.64	0.02336
After pregnancy	1.05	0.93	1.20	0.42067
Assault PEDW (mother)				
	1.07	0.87	1.32	0.53618
Diabetes PEDW (child)				
	1.43	0.54	3.79	0.46615
Diabetes GP (child)				
	0.58	0.21	1.60	0.29455
Epilepsy PEDW (child)				
	2.09	1.62	2.71	0.00000
Epilepsy GP (child)				
	2.95	2.39	3.66	0.00000
Asthma PEDW (child)				
	1.12	1.03	1.22	0.00837
Asthma GP (child)				
	0.91	0.88	0.95	0.00000
Ear PEDW (child)				
	1.36	1.26	1.45	0.00000
Eye PEDW (child)				
	1.30	1.18	1.42	0.00000
Any emergency hospital admission (child)				
	1.09	1.05	1.13	1.09
Any A&E attendance (child)				

(*Continued*)

Table 2. (Continued)

Variable name	OR	Lower CI	Upper CI	P value
	1.02	0.98	1.06	1.02
Free school meal				
	1.85	1.78	1.93	0.00000
Local authority				
Blaenau Gwent	1.00	0.87	1.14	0.95948
Bridgend	0.86	0.75	0.98	0.02302
Caerphilly	1.14	1.01	1.28	0.04063
Cardiff	1			
Carmarthenshire	1.28	1.12	1.47	0.00032
Ceredigion	1.13	0.94	1.35	0.19523
Conwy	1.70	1.47	1.97	0.00000
Denbighshire	1.10	0.93	1.28	0.25936
Flintshire	1.27	1.10	1.46	0.00095
Gwynedd	1.26	1.09	1.46	0.00162
Isle of Anglesey	1.13	0.97	1.31	0.12881
Merthyr Tydfil	1.09	0.94	1.27	0.26314
Monmouthshire	1.08	0.89	1.30	0.44402
Neath Port Talbot	1.62	1.43	1.84	0.00000
Newport	0.79	0.69	0.91	0.00118
Pembrokeshire	1.12	0.95	1.33	0.17595
Powys	1.05	0.88	1.25	0.59866
Rhondda Cynon Taff	1.23	1.09	1.39	0.00078
Swansea	1.49	1.32	1.68	0.00000
Torfaen	0.96	0.83	1.12	0.62127
Vale of Glamorgan	1.02	0.88	1.17	0.81997
Wrexham	1.22	1.06	1.39	0.00419
WIMD 2014—overallf				
1 (most deprived)	1.67	1.57	1.77	0.00000
2	1.52	1.43	1.62	0.00000
3	1.37	1.29	1.46	0.00000
4	1.28	1.19	1.37	0.00000
Local area—urban/rural				
Rural town	1.08	1.03	1.13	0.00277
Rural village	1.20	1.12	1.29	0.00000
No of adult in the household				
1	1.08	1.04	1.12	0.00028
3	1.16	1.10	1.22	0.00000
4 or above	1.17	1.10	1.24	0.00000
No of children in the household (excluding the cohort member)				
0	1.08	1.03	1.13	0.00220
2	1.21	1.16	1.26	0.00000
3	1.43	1.35	1.52	0.00000
4 or above	1.63	1.52	1.75	0.00000
Child's age in the academic year				
	0.31	0.30	0.33	0.00000
School session absences				
	1.02	1.02	1.02	0.00000

(*Continued*)

Table 2. (Continued)

Variable name	OR	Lower CI	Upper CI	P value
Unauthorised absences				
	1.00	0.99	1.00	0.00064

https://doi.org/10.1371/journal.pone.0273596.t002

The findings from our study suggest that rising poverty and the cost-of-living crisis are likely to result in lower school readiness and lower educational attainment. This will put a strain on school resources as children enter school [31]. Children in family and area-level deprivation are at higher risk of not being school ready. This finding is consistent with the existing literature [7,15]. Boys being disadvantaged compared to girls has been noted in other research [32]. In fact, it is suggested that family instability (separation, divorce, second families) affects boys more than girls, with a lack of a male influence impacting on behavioural difficulties [32] and that recent population increases in family instability can help explain a trend in lower attainment for boys at all levels. In addition, existing research clearly demonstrates that deprivation is a strong predictor of low school readiness [9,33]. Various indicators of deprivation such as parental employment, lower parental educational attainment, lower income, less time with the child, poorer play/local area facilities have been identified as significantly linked with low school readiness [7]. Our findings such as the significant association between living in family level (eligibility for FSM) and area level (most deprived WIMD) deprivation and higher chance not to be school ready are along the similar lines reported in the literature [9,34,35]. Hence, it is suggested that pre-school investment [35] and free childcare can overcome some of the risk factors associated with deprivation.

In this study, the decision tree model highlighted the risk factors which are clustered together e.g., boys living in household level deprivation and higher absences in school are at high risk of low school readiness. Similarly, girls who are often missing school are at risk of not being school ready. It also showed that children who are not breastfed, having ear infection and younger in academic year than their peers will more likely be not school ready. The branches of clustered risk factors are used to examine the determinants of low school readiness. These most significant risk factors can contribute to understand the profile of the

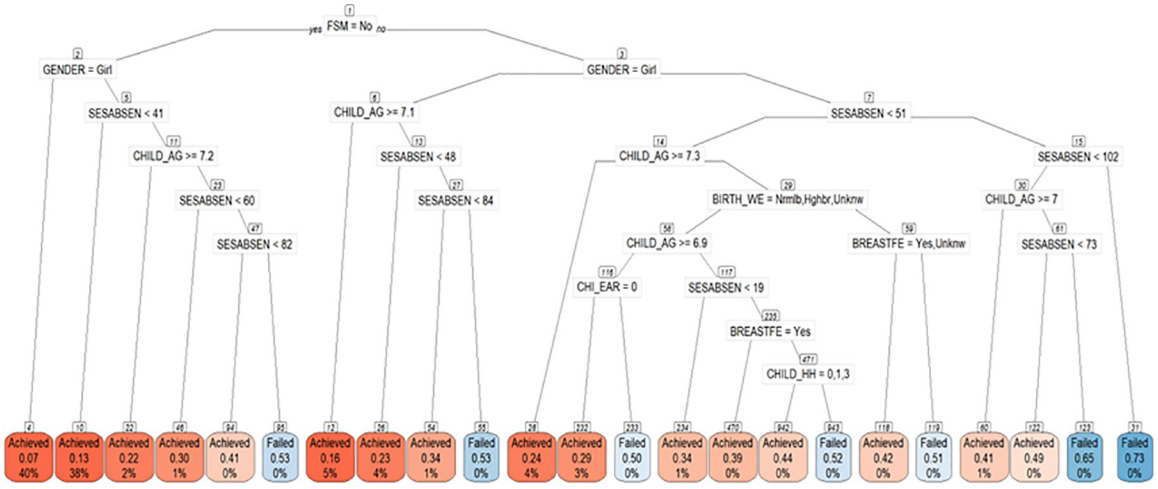


Fig 2. Decision tree for the children without learning difficulty.

https://doi.org/10.1371/journal.pone.0273596.g002

Table 3. Confusion matrix/two by two table of the DT model.

Prediction	Reference (Children without learning difficulty) n = 14,575	
	Did not achieve (P)	Achieved (N)
Did not achieve(P)	108 (TP)	78 (FP)
Achieved (N)	2,078 (FN)	12,311 (TN)

https://doi.org/10.1371/journal.pone.0273596.t003

vulnerable children and their families and help to improve the decision making at policy level which will support children to overcome the odds and have a better start in life.

This work has been developed as part of the Early Years Vulnerability Profiling Pilot. This will enable the Health Board and local authority to plan how the Early Years Vulnerability Profile can be used to inform better targeting of prevention and early intervention to children and their families up to the age of seven years to enable better outcomes for health, well-being, education, and social skills. The clustered risk factors can be used to understand what is associated with as determinants of low school readiness at population level. This can contribute to informed decision making at policy level that supports the children to have best start in life.

Strengths and limitations

This study is based on linked data for an entire country over a 6-year period. This provides a wider range of risk factors from routine administrative data at a national level which can be addressed to improve outcomes for children who are exposed to inequalities and disadvantages from early life. It can contribute to breaking a cycle of disadvantage for children by helping to identify where and how to target early years interventions designed to improve school readiness. There is evidence that routinely collected data observed during perinatal period can contribute to improve child's development at early years [15,21]. A linked population level database can facilitate a holistic investigation of the complex factors associated with the low school readiness. Longitudinal data linkage allows the capturing of the developmental trajectory of the individual child from school foundation phase and health visitor records. Combing this with maternal physical and mental health records can only strengthen the power of the analysis, as it is proved that maternal health and wellbeing is one of the biggest predictors of child's development and wellbeing (*Improving school readiness Creating a better start for London*). If all these information can be available at an early stage to the policy makers from school and health visitors report, this can directly contribute to identify the most vulnerable children and their families at a very early stage and can help to build necessary intervention and support plans for them when it's most needed.

However, it can only examine factors which are recorded using routine data. Important factors such as parenting style, time spent with the child reading, playing, and interacting cannot be captured with this data but would be important factors associated with school readiness. Another limitation of the study is that it only included the children who were in Wales during the entire study period and were removed if the children moved out of Wales as we were unable to capture their exposure records. However, this would not lead to any selection issues (please

Table 4. Prediction model performance (n = 14,575 children from Rhondda Cynon Taff).

	Accuracy	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Prevalence
Decision Tree	85.21%	4.94%	99.37%	58.06%	85.56%	15%

https://doi.org/10.1371/journal.pone.0273596.t004

see Appendix 5 in S1 File) since these were arbitrary and independent events and does not lead to or is not linked with low school readiness. The study has identified a cluster of socioeconomic, health and household level risk factors leading to low school readiness and establishing a direct causal pathway of the modifiable risk factors is beyond the scope of the study.

A major strength of the study is that it incorporated data from birth till they enter their formal school to build the model to identify the risk factors of low school readiness, hence these findings can be helpful to identify the children at risk of low school readiness before they start their schooling as many of these factors are present in the first years of life (gender, deprivation, gestational age, parental health) and so those at risk can be supported through access to childcare, parenting support and supporting breastfeeding. In addition, the school readiness for local children coming to a school can be predicted and this means schools can have the necessary resources in place to help the specific catchment of children coming to their school.

Conclusion

This study highlighted a vulnerability profile of the children who are at higher risk of low school readiness by identifying the group of risk factors which are clustered together. The findings suggest that earlier intervention (access to childcare, mother/baby groups, community activities, parenting interventions) could help to improve the outcomes for children who are at a high risk of low school readiness. This is especially true in deprived areas with low access to childcare and where there are child or adult health problems. It has been observed that intervention programmes like Flying Start has positive effects on the children living in deprivation including improved school attendance and better educational outcomes than their peers who are in similar condition but not under Flying Start programme [36]. This work suggests that interventions which focused on boys in deprived areas, encourage or facilitated attendance in nursery in the early years, investment in early years childcare and promoting breastfeeding would have a significant impact on school readiness. Interventions such as parenting programmes which supported families with parental learning difficulties, support when there is parental or child illness (e.g., community tutoring volunteer programmes) especially for epilepsy would make a significant difference for the child's readiness for school. This could positively influence a child's life trajectory by strengthening foundations for lifelong learning, improving health and wellbeing outcomes throughout the life-course, and reducing education and developmental inequalities that persist.

Supporting information

S1 File.
(ZIP)

Acknowledgments

This work was also supported by the National Centre for Population Health and Well-Being Research (NCPHWR) which is funded by Health and Care Research Wales. This work was supported by Health Data Research UK which receives its funding from HDR UK Ltd (NIWA1) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Welcome Trust.

This work uses data provided by patients and collected by the NHS as part of their care and support. Anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank has been used in this study. We would like to acknowledge all the data providers who enable SAIL to make anonymised data available for research.

Author Contributions

Conceptualization: Amrita Bandyopadhyay, Emily Marchant, Julie Evans, Sinead Brophy.

Data curation: Amrita Bandyopadhyay.

Formal analysis: Amrita Bandyopadhyay, Hope Jones, Michael Parker.

Funding acquisition: Julie Evans, Sinead Brophy.

Investigation: Amrita Bandyopadhyay, Emily Marchant.

Methodology: Amrita Bandyopadhyay, Sinead Brophy.

Project administration: Amrita Bandyopadhyay.

Resources: Amrita Bandyopadhyay, Emily Marchant.

Software: Amrita Bandyopadhyay.

Supervision: Sinead Brophy.

Validation: Amrita Bandyopadhyay, Hope Jones, Michael Parker.

Visualization: Amrita Bandyopadhyay.

Writing – original draft: Amrita Bandyopadhyay.

Writing – review & editing: Amrita Bandyopadhyay, Emily Marchant, Hope Jones, Michael Parker, Julie Evans, Sinead Brophy.

References

1. Anderson LM, Shinn C, Fullilove MT, Scrimshaw SC, Fielding JE, Normand J, et al. The effectiveness of early childhood development programs: A systematic review. Am J Prev Med. 2003; 24: 32–46. https://doi.org/10.1016/S0749-3797(02)00655-4 PMID: 12668197
2. Burger K. How does early childhood care and education affect cognitive development? An international review of the effects of early interventions for children from different social backgrounds. Early Child Res Q. 2010; 25: 140–165. https://doi.org/10.1016/j.ecresq.2009.11.001
3. Duncan GJ, Dowsett CJ, Claessens A, Magnuson K, Huston AC, Klebanov P, et al. School readiness and later achievement. Dev Psychol. 2007; 43: 1428. https://doi.org/10.1037/0012-1649.43.6.1428 PMID: 18020822
4. High PC. School Readiness. Pediatrics. 2008; 121: e1008–e1015. https://doi.org/10.1542/peds.2008-0079 PMID: 18381499
5. Lipscomb ST, Miao AJ, Finders JK, Hatfield B, Kothari BH, Pears K. Community-Level Social Determinants and Children's School Readiness. Prev Sci. 2019; 20: 468–477. https://doi.org/10.1007/s11121-019-01002-8 PMID: 30852712
6. Hammer CS, Morgan P, Farkas G, Hillemeier M, Bitetti D, Maczuga S. Late Talkers: A Population-Based Study of Risk Factors and School Readiness Consequences. J Speech Lang Hear Res JSLHR. 2017; 60: 607–626. https://doi.org/10.1044/2016_JSLHR-L-15-0417 PMID: 28257586
7. Camacho C, Straatmann VS, Day JC, Taylor-Robinson D. Development of a predictive risk model for school readiness at age 3 years using the UK Millennium Cohort Study. BMJ Open. 2019; 9: e024851. https://doi.org/10.1136/bmjopen-2018-024851 PMID: 31213442
8. Hair E, Halle T, Terry-Humen E, Lavelle B, Calkins J. Children's school readiness in the ECLS-K: Predictions to academic, health, and social outcomes in first grade. Early Child Res Q. 2006; 21: 431–454. https://doi.org/10.1016/j.ecresq.2006.09.005

9. Kiernan KE, Mensah FK. Poverty, Maternal Depression, Family Status and Children's Cognitive and Behavioural Development in Early Childhood: A Longitudinal Study. J Soc Policy. 2009; 569–588. Available: https://doi.org/10.1017/S0047279409003250.
10. Hobcraft J, Kiernan KE. Predictive factors from age 3 and infancy for poor child outcomes at age 5 relating to children's development, behaviour and health: evidence from the Millennium Cohort Study. Univ York York. 2010.
11. Nelson J, Martin K, Featherstone G. What Works in Supporting Children and Young People to Overcome Persistent Poverty?: A Review of UK and International Literature. National Foundation for Educational Research; 2013.
12. PhD MJ, Duku E. The School Entry Gap: Socioeconomic, Family, and Health Factors Associated With Children's School Readiness to Learn. Early Educ Dev. 2007; 18: 375–403. https://doi.org/10.1080/10409280701610796a
13. Shah PS. Paternal factors and low birthweight, preterm, and small for gestational age births: a systematic review. Am J Obstet Gynecol. 2010; 202: 103–123. https://doi.org/10.1016/j.ajog.2009.08.026 PMID: 20113689
14. Kelly Y, Sacker A, Gray R, Kelly J, Wolke D, Quigley MA. Light drinking in pregnancy, a risk for behavioural problems and cognitive deficits at 3 years of age? Int J Epidemiol. 2009; 38: 129–140. https://doi.org/10.1093/ije/dyn230 PMID: 18974425
15. Chittleborough CR, Searle AK, Smithers LG, Brinkman S, Lynch JW. How well can poor child development be predicted from early life characteristics?: A whole-of-population data linkage study. Early Child Res Q. 2016; 35: 19–30. https://doi.org/10.1016/j.ecresq.2015.10.006
16. Barnett WS, Belfield CR. Early Childhood Development and Social Mobility. Future Child. 2006; 16: 73–98. Available: https://doi.org/10.1353/foc.2006.0011 PMID: 17036547
17. Department for Education. GCSE and equivalent attainment by pupil characteristics: 2012. 2013 Jan. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/219337/sfr04-2013.pdf.
18. Clifton J, Cook W. A long division: Closing the attainment gap in England's secondary schools. Lond IPPR. 2012.
19. Black M, Barnes A, Baxter S, Beynon C, Clowes M, Dallat M, et al. Learning across the UK: a review of public health systems and policy approaches to early child development since political devolution. J Public Health. 2020; 42: 224–238. https://doi.org/10.1093/pubmed/fdz012 PMID: 30799501
20. Lynch JW, Law C, Brinkman S, Chittleborough C, Sawyer M. Inequalities in child healthy development: Some challenges for effective implementation. Soc Sci Med. 2010; 71: 1244–1248. Available: https://ideas.repec.org//a/eee/socmed/v71y2010i7p1244-1248.html. https://doi.org/10.1016/j.socscimed.2010.07.008 PMID: 20691527
21. Brinkman SA, Gregory TA, Goldfeld S, Lynch JW, Hardy M. Data Resource Profile: The Australian Early Development Index (AEDI). Int J Epidemiol. 2014; 43: 1089–1096. https://doi.org/10.1093/ije/dyu085 PMID: 24771275
22. Brinkman S, McDermott R, Lynch J. Better understanding trajectories of child development: opportunities for data linkage with the Australian Early Development Index (AEDI). Public Health Bull S Aust. 2010; 7: 7–10. Available: https://researchonline.jcu.edu.au/36811/.
23. Department for Education and Skills. Curriculum for Wales: Foundation phase framework. Welsh Government Cardiff; 2015. Available: https://hwb.gov.wales/storage/d5d8e39c-b534-40cb-a3f5-7e2e126d8077/foundation-phase-framework.pdf.
24. Ford DV, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Serv Res. 2009; 9: 157. https://doi.org/10.1186/1472-6963-9-157 PMID: 19732426
25. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke J-P, Ford DV, et al. The SAIL databank: linking multiple health and social care datasets. BMC Med Inform Decis Mak. 2009; 9: 3. https://doi.org/10.1186/1472-6947-9-3 PMID: 19149883
26. Rodgers SE, Lyons RA, Dsilva R, Jones KH, Brooks CJ, Ford DV, et al. Residential Anonymous Linking Fields (RALFs): a novel information infrastructure to study the interaction between the environment and individuals' health. J Public Health. 2009; 31: 582–588. https://doi.org/10.1093/pubmed/fdp041 PMID: 19447812
27. Welsh Government. Welsh Index of Multiple Deprivation (full Index update with ranks): 2011. In: GOV.WALES [Internet]. 31 Aug 2011 [cited 5 Aug 2020]. Available: https://gov.wales/welsh-index-multiple-deprivation-full-index-update-ranks-2011.
28. Hughes S. Foundation Phase Outcomes and National Curriculum Teacher Assessment of Core Subjects at Key Stages 2 and 3, 2018. Welsh Government; 2018. Available: https://gov.wales/sites/default/

files/statistics-and-research/2018-12/180808-foundation-phase-outcomes-national-curriculum-teacher-assessment-core-subjects-key-stages-2-3-2018-en.pdf.

29. Atkinson Beth. rpart function | R Documentation. [cited 14 Jan 2021]. Available: https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart.
30. Lewis RJ. An introduction to classification and regression tree (CART) analysis. Annual meeting of the society for academic emergency medicine in San Francisco, California. 2000.
31. Lloyd-Newman E. New School Readiness Report Released. Kindred^2; 2023 Jan. Available: https://kindredsquared.org.uk/wp-content/uploads/2023/01/Kindred-Squared-School-Readiness-Report.pdf.
32. Cooper CE, Osborne CA, Beck AN, McLanahan SS. Partnership Instability, School Readiness, and Gender Disparities. Sociol Educ. 2011; 84: 246–259. https://doi.org/10.1177/0038040711402361 PMID: 21949448
33. improving_school_readiness.pdf. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/781623/improving_school_readiness.pdf.
34. Ryan RM, Fauth RC, Brooks-Gunn J. Childhood Poverty: Implications for School Readiness and Early Childhood Education. Handbook of research on the education of young children, 2nd ed. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; 2006. pp. 323–346.
35. Ferguson H, Bovaird S, Mueller M. The impact of poverty on educational outcomes for children. Paediatr Child Health. 2007; 12: 701–706. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2528798/. https://doi.org/10.1093/pch/12.8.701 PMID: 19030450
36. Wilton J, Government W, Davies R. Flying Start Evaluation: Educational Outcomes.: 76.

My input

In this paper, I developed a comprehensive research plan, which included conducting a literature review to determine the covariates for low school readiness, derived from numerous routine administrative datasets. I performed extensive data linkage and harmonisation across these datasets to prepare the final dataset for analysis. Utilising the R software package and SQL, I constructed the LR and DT models. Beyond data preparation and analysis, I also authored and published this journal article as both the first author and corresponding author.

Impact

- This article has been published in *PLOS One* in 2023.
- The paper has been cited in four other published works.
- The findings have also been utilised as the outcome of the pilot program '*Vulnerability Profiling*' conducted by Public Health Wales, serving as the foundation for their future research and initiatives
- The findings of the school readiness paper have been reported as a Data Insight report across the UK.

Conclusion

School readiness is a critical milestone in a child's developmental trajectory, and disparities in preparedness at school entry can have long-lasting effects on academic performance and life opportunities. This chapter highlights the key role of socio-economic conditions, early health status and early-life interventions in shaping educational outcomes. The findings suggest that policies aimed at reducing inequalities in early-life education and providing targeted support to children at risk of low school readiness could help mitigate long-term disparities in learning and development. Educational preparedness is one aspect of childhood vulnerability. The next chapter shifts focus to understanding how local area deprivation influences life chances for children growing up in poverty in Wales. By integrating health, education and social care data, this study provides insight into the role of community-level factors in shaping child outcomes and resilience.

Chapter 4: How does the local area deprivation influence life chances for children in poverty in Wales: A record linkage cohort study

Critical summary

Background

Living in poverty, both at the family and area-levels, has a profoundly detrimental impact on a child's development. Townsend et al, in their study has established a clear correlation between health inequalities within the population and regional disparities in socio-economic conditions in North of England (71). Their findings also emphasises the detrimental effect of regional disparities in wealth on the health outcome of the population. A systematic review conducted by Visser K et al., highlighted the effects of neighbourhood deprivation on young people's wellbeing and mental health (72). They primarily included 30 studies (including UK, USA and Europe) in their review which measured deprivation in the form of neighbourhood socio-economic indicators (neighbourhood average/median income, employment rates, educational levels). However, they also highlighted the fact that these effects remain largely underexplored. In Wales, the influence of local area deprivation on child development has not been thoroughly investigated. This study, titled 'How Does Local Area Deprivation Influence Life Chances for Children in Poverty in Wales: A Record Linkage Cohort Study,' is the first of its kind to examine how local area conditions affect children living in poverty in Wales and contribute to their strength to overcome the odds using solely routine data.

Utilisation of administrative data

This longitudinal record linkage study combined routine administrative datasets, including Pre 16 Education data in Wales, WDS (a Wales-wide administrative register for all individuals with a general practitioner), WLGP (primary care records in the Welsh Longitudinal General Practice) and PEDW (secondary care health records) within the SAIL Databank platform. For this study, a nationally representative cohort of children was constructed using the anonymised and encrypted person-based identifier ALF, enabling follow-up through to their Key Stage 4 (KS4) attainments. This research utilised free school meal (FSM) eligibility as a proxy measure of family level disadvantage, derived from routine data (73). Additionally, the Welsh Index of Multiple Deprivation (WIMD) 2011 (74) was used as an indicator of local area-level deprivation. The study successfully developed an outcome variable known as the Profile to Leave Poverty (PLP), which serves as an indicator of a child's likelihood of overcoming the adverse outcomes associated with poverty. This resilience profile for children was constructed using four key markers: a) achieving KS4, b) absence of a mental health condition, c) no substance misuse and

d) no record of alcohol abuse in routine healthcare data. This study investigated how, despite growing up in poverty, these children developed a resilience profile characterised by educational attainment, no development of mental health conditions and no signs of risk-taking behaviours.

In this study, the population comprised of children who completed their age 16 exams (Key Stage 4 (KS4)) between 2009 and 2016 and had a valid Free School Meal (FSM) record (binary variable) available in the Pre-16 Education data in Wales. The children who did not have a continuous residential record in Wales between the ages of six months and KS4 exams were removed from the study to ensure complete coverage of the data. The outcome data (PLP), including mental health records, alcohol records and substance misuse data, were obtained for study participants aged between 11 and KS4. The exposure data including the residential records (including WIMD) were obtained when the children completed their KS4 exams. Previous house moves were not considered in this study due to the complex nature of the data, which has been acknowledged as a limitation of the study.

The novelty of this study lies in its ability to utilise linked routine administrative data to create a comprehensive framework for investigating the area-level impact on children in poverty within a nationally representative population in Wales. This approach not only enhances the robustness of the findings but also provides a nuanced understanding of the characteristics of supportive neighbourhoods that enable children to overcome vulnerabilities. My work also emphasises the reusability of the data, as the database and the resilience variable are available inside SAIL for future use.

Application of data science methods

In this paper, stepwise LR was implemented to evaluate the area-level factors contributing to children's resilience, particularly among those living in poverty. This methodological approach facilitated the selection of the most relevant covariates associated with the outcome variable, PLP, based on their statistical significance, employing a data-driven model that strengthens the robustness of the findings. The stepwise LR model effectively managed multiple predictors while controlling for confounding variables. This method enabled the assessment of the relationships between local area deprivation and its various components, including income, employment, health, environment, safety and access to neighbourhood services. Additionally, it examined their impact on child outcomes, such as mental health, academic achievement and substance misuse. By employing this rigorous analytical approach, the study enhances the validity of the findings, providing a clearer understanding of how different factors interact and contribute to a child's resilience. The finding of this work highlights the importance of adopting a multifaceted perspective on deprivation when developing interventions aimed at improving the life chances of children in poverty.

Early-life vulnerability profiling

This research aimed to make a significant contribution to the field of early-life vulnerability profiling by investigating the complex dynamics between local area characteristics and a child's overall outcomes. This research question has been one of the priorities of the Welsh Government's Early Years programme, Administrative Data Research (ADR), Wales, and understanding this dynamics is essential for developing targeted interventions that can mitigate the adverse effects of poverty and promote equitable opportunities for all children. The study's important findings highlighted specific neighbourhood characteristics, such as community safety, area income and accessibility to local services. These factors play a crucial role in helping children overcome the adverse impacts of poverty. For instance, neighbourhoods with higher levels of community safety provide a secure environment that fosters emotional wellbeing and stability, allowing children to thrive. Furthermore, areas with higher income levels often offer better educational resources and opportunities, which can significantly enhance children's academic performance and social development. The study also highlighted the necessity of better connectivity within communities. Children with access to supportive community resources, such as healthcare, recreational facilities and mentorship programs, are better equipped to navigate challenges and build resilience. By leveraging existing data, this research provides valuable insights into the factors that contribute to resilience in children facing adversity, informing targeted interventions and policy decisions aimed at improving life chances for disadvantaged youth.

Published journal paper

SSM - Population Health 22 (2023) 101370

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: www.elsevier.com/locate/ssmph

How does the local area deprivation influence life chances for children in poverty in Wales: A record linkage cohort study

Amrita Bandyopadhyay [a,b,*], Tony Whiffen [c], Richard Fry [b,d], Sinead Brophy [a,b,d]

[a] National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Wales, SA2 8PP, UK
[b] Administrative Data Research Unit, Swansea University, Wales, SA2 8PP, UK
[c] Administrative Data Research Unit, Welsh Government, Wales, CF10CF10 3NQ, UK
[d] Health Data Research UK, Swansea University Medical School, Wales, SA2 8PP, UK

ARTICLE INFO

Keywords:
Local area
Deprivation
Child poverty
Resilience
Education
Record linkage
Cohort study

ABSTRACT

Objectives: Children growing up in poverty are less likely to achieve in school and more likely to experience mental health problems. This study examined factors in the local area that can help a child overcome the negative impact of poverty.
Design: A longitudinal record linkage retrospective cohort study.
Participants: This study included 159,131 children who lived in Wales and completed their age 16 exams (Key Stage 4 (KS4)) between 2009 and 2016. Free School Meal (FSM) provision was used as an indicator of household-level deprivation. Area-level deprivation was measured using the Welsh Index of Multiple Deprivation (WIMD) 2011. An encrypted unique Anonymous Linking Field was used to link the children with their health- and educational records.
Outcome measures: The outcome variable 'Profile to Leave Poverty' (PLP) was constructed based on successful completion of age 16 exams, no mental health condition, no substance and alcohol misuse records in routine data. Logistic regression with stepwise model selection was used to investigate the association between local area deprivation and the outcome variable.
Results: 22% of children on FSM achieved PLP compared to 54.9% of non-FSM children. FSM Children from least deprived areas were significantly more likely to achieve PLP (adjusted odds ratio (aOR) - 2.20 (1.93, 2.51)) than FSM children from most deprived areas. FSM children, living in areas with higher community safety, higher relative income, higher access to services, were more likely to achieve PLP than their peers.
Conclusion: The findings indicate that community-level improvements such as increasing safety, connectivity and employment might help in child's education attainment, mental health and reduce risk taking behaviours.

1. Introduction

Latest figures suggest that in 2020, 29.3% of children aged between 0 and 19 are living in poverty (i.e. family income below 60% of the median income) in Wales, which is a 1% rise compare to the previous year (Observatory, n.d.). Living in persistent poverty has a detrimental impact on child health, cognitive and behavioural outcomes (Wickham et al., 2016). Child poverty has caused an unprecedented increase in infant mortality in recent years in the UK (Taylor-Robinson et al., 2019). After a steady fall in the last decade (post-2010), the child poverty rate has also now started to increase in the UK (Joyce, 2014; Taylor-Robinson et al., 2019). In the post-recession recovery period (i.e. since 2008) inequality increased because of disproportionately slow recovery for low-income families (Beatty & Fothergill, 2016; Cribb et al., 2018). This is due to real-term cuts in benefits, increasing housing costs and restricted possibilities to improve income from work (e.g. due to salary reductions, freeze in promotions) (Lambie-Mumford & Green, 2017). As a result, of all children living in relative poverty, the majority are from working families as opposed to workless households (Vizard et al., 2019). Currently 67% of the children in relative poverty are living in households where at least one person is working (Welsh Government, 2019b). A report from *End Child Poverty* carried out by Loughborough University has shown that child poverty is disproportionately rising in the UK's most impoverished areas (Loughborough, 2019). The report

* Corresponding author. National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Wales, SA2 8PP, UK.
E-mail address: [redacted] (A. Bandyopadhyay).

https://doi.org/10.1016/j.ssmph.2023.101370
Received 3 October 2022; Received in revised form 16 February 2023; Accepted 18 February 2023
Available online 23 February 2023

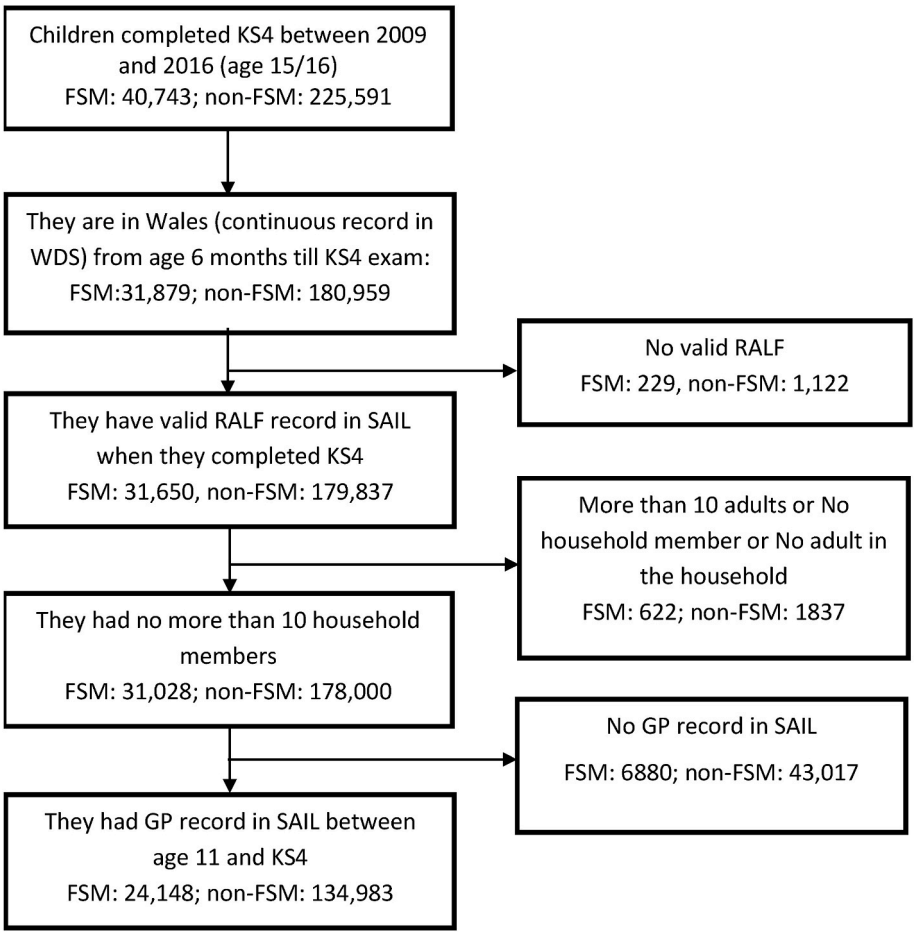


Fig. 1. Participants consort diagram (based on Free School Meal eligibility).

shows that in some parts of Wales, children from deprived families are six times more likely to grow up in poverty than their neighbours if they are living in less deprived areas. The latest report from the Welsh Index of Multiple Deprivation (WIMD) 2019 from Welsh Government highlighted 'deep-rooted' deprivation by highlighting the areas in Wales which have remained as the top most deprived areas for more than last 15 years, which indicates a lack of social mobility in most of these areas (Welsh Government, 2019a).

A child growing up in a deprived area implies that they are more likely to be provided with insufficient educational support, lack of recreational space (no safe park or playground) and receive poor quality childcare and health support (Galster et al., 2007). This has numerous inevitable long-term consequences such as poorer mental and physical health, lower school achievement, and worse outcomes in adulthood (Featherstone et al., 2019; Galster et al., 2007; Wickham et al., 2016; Wood, 2003). Another study has found that the children in deprived area are at higher risk of early alcohol use (Bandyopadhyay et al., 2022) and early onset of alcohol use increases the risk of alcohol dependence and other illicit drug use in later life (Hingson et al., 2006). Children living in deprived neighbourhoods are less likely to complete high school and achieve higher educational attainment. This creates a significant difference in their earning levels in later life compared to their peers (Galster et al., 2007). Local areas with community safety issues often restrict children from after-school outdoor activities and increases their sedentary behaviours. This significantly contributes to childhood obesity amongst children living in poor neighbourhoods (Cecil-Karb & Grogan-Kaylor, 2009). Family- and area level disadvantageous socio-economic conditions often lead to teenage pregnancy (Penman-Aguilar et al., 2013), which is significantly associated with adverse health outcomes and social consequences (Cook & Cameron, 2017).

Although growing up in a deprived family and local area increase the risk of adverse consequences in their life, some children are more able to beat the odds than their peers despite coming from disadvantage, and show resilience (Sattler & Gershoff, 2019). Studies have investigated various factors that can be linked with overcoming odds, such as moving to a more affluent area in early childhood (Chetty et al., 2016), living in an area with better access to green space (Flouri et al., 2014), safer community areas so that parents allow and encourage their children to be involved in outdoor physical activity (Veitch et al., 2013), and neighbourhood safety that enhances collective socialisation (Marco & Vernon-Feagans, 2013; Minh et al., 2017). Though such evidence is fragmented, it indicates that improvement of the quality, facility and environment of the local area can help the children to build resilience and overcome adversity. Hence it is necessary to develop a holistic understanding of neighbourhoods and prioritise the various aspects of a local area which can help children and their parents to improve their life and overcome poverty.

Family level deprivation as a key indicator of child's poor development has been discussed in literature, but this study focuses on whether a local area level improvement can moderate this relationship. This study investigates the socio-economic determinants of a local area that are associated with the resilience in children using a linked routine data framework. The aim of the study is to develop a holistic understanding of various aspects of a local area which contribute to the resilience of the children and can help children to improve their life. This study has used the deprivation index WIMD 2011 to identify concentrations and variations of several domains of deprivation for small areas in Wales and its impact on children (n = 159,131). This work has developed a profile of

children showing resilience despite family level deprivation based on factors which have significant association with improving their lives and overcoming poverty. This has been modelled as 'Profile to Leave Poverty' (PLP) and it has been derived based on four major components (education, mental health, alcohol, and substance misuse). The findings of the study can provide important insights for targeted policy development and intervention.

2. Method

2.1. Sample

The study population was comprised of children who completed their age 16 exams (Key Stage 4 (KS4)), between 2009 and 2016 and had a valid Free School Meal (FSM) record (eligible or not eligible). The selected children were either born or resident of Wales until they completed KS4. The participants were derived by linking Wales Demographic Service Dataset (WDSD) (a Wales-wide administrative register for all individuals with a general practitioner (GP)) and education datasets. Data linkage was performed with the help of an anonymised encrypted linkage key known as Anonymous Linking Field (ALF) provided by trusted third party in the Secure Anonymous Information Linkage (SAIL) databank platform at Swansea University (Ford et al., 2009; Lyons et al., 2009). To enable individuals living in the same household to be anonymously linked, Residential Anonymous Linking Field (RALFs) were created by encrypting individual's address identifiers for the study period (Johnson et al., 2021). The children who did not have a continuous residential record (valid RALF) in WDSD between age six months and KS4 exam (to ensure they lived in Wales throughout the childhood, and we had valid measures of exposures) and primary care record in Welsh Longitudinal General Practice (WLGP) dataset between age 11 and KS4 (when the outcome variable was observed) were excluded from the analysis to ensure the complete data coverage and follow-up period. A detailed participants flow diagram of the study population is provided in Fig. 1.

2.2. Exposure variables

In this study local area deprivation was measured by using WIMD 2011 (Welsh Government, 2011) which is the official measure of relative deprivation for small areas in Wales. Lower layer Super Output Areas (LSOA) are the geographic units used to define small areas in Wales and England. There were 1896 LSOAs in Wales and WIMD 2011 ranked all LSOAs (1 most deprived to 1896 least deprived). The study used WIMD 2011 as this was timely with the study period. In this study, individual's residential identifier RALF was linked to LSOAs, which are linked with WIMD rank aggregated into a quintile scale where a lower value denotes greater deprivation. Considering the statistical significance of the categories with respect to the study population and the interpretability of the findings, the study considered WIMD aggregated into a quintile scale instead of by decile. Along with overall WIMD rank, component scores for WIMD domains such as income, community safety, health, access to services, physical environment, housing, but not education have been considered as main exposure variables. For individuals, household level deprivation has been measured by FSM eligibility at KS4 (Taylor, 2018).

2.3. Covariates

The other covariates that were included in the study are - living in urban or rural area, number of adults and number of children in the household, living with someone who had depression (diagnosis and/or medication), any household member diagnosed with serious mental illness such as schizophrenia, bipolar disorder (for ICD10 and Read codes see Supplementary material Codes 1), household member who had an alcohol related hospitalisation record (for ICD10 codes see Supplementary material Codes 2) and whether the child needs special education support. Since the study builds a cohort of children who are completing KS4 between 2009 and 2016, hence to adjust the effect of different academic years, their KS4 assessment year (Exam Year) has been considered in the analysis.

2.4. Outcome variable

The study aimed to build a profile that can contribute to the resilience of the children. 'Profile to Leave Poverty' (PLP) is an indicator of overcoming poverty at the transition between adolescent and early adulthood. The resilience profile of the children known (PLP) has been developed with the four major components such as: a) poor educational attainment, b) developing mental health condition, c) early alcohol use, and d) early substance misuse. The existing literature has already shown the significant association between poverty and these four major components. It has been identified that the children living poverty are more likely to be affected by these four risk factors which will have several detrimental impacts on their later life (this has been discussed in the introduction section). Hence, the study developed a resilience profile by adding all four components where there are positive outcome from all four factors. The PLP has been derived based on the following four criterions –

a. Achieved KS4: If they have successfully completed L2EWM (level 2 English/Welsh Maths– A* to C in 5 GCSE subjects including Maths and English/Welsh)
b. No mental health condition: They have no records of the following conditions - Attention Deficit Hyperactive Disorder, Conduct Disorder, Depression, Serious Mental Illness, Self-harm between age 11 and KS4 assessment
c. No substance misuse: They have no substance misuse record between age 11 and KS4 assessment
d. No alcohol abuse: They have no alcohol related records between age 11 and KS4 assessment

The children who satisfied all four above-mentioned conditions were considered as 'achieved' PLP. Those who did not satisfy one of the conditions were considered as 'not achieved' PLP, i.e.

PLP 'achieved' = KS4 achieved *AND* No mental health condition record *AND* No substance misuse record *AND* No alcohol abuse record

PLP 'not achieved' = KS4 not achieved *OR* mental health condition record *OR* substance misuse record *OR* alcohol abuse record

The study population has been linked with relevant education data to obtain the KS4 record. Mental health, substance misuse and alcohol records were derived from hospital admissions dataset known as Patient Episode database in Wales (PEDW), primary care dataset known Welsh Longitudinal General Practice (WLGP) and substance misuse dataset. ICD-10 codes used in PEDW indicate hospital admission due to mental health conditions, substance misuse and alcohol whilst GP-recorded Read codes highlight diagnosis and medication associated with mental health conditions, substance misuse and alcohol in primary care health system. ICD-10 and Read codes are mentioned the Supplementary material Codes 3, 4 & 5.

2.5. Statistical analysis

The study primarily aimed to investigate the association between the resilience profile PLP derived by the study and the local area deprivation measured by WIMD among the children living in high household-level deprivation (FSM children). The current study, however, also investigated a similar association among the non-FSM children group, hence FSM-stratified analysis was performed. A supplementary analysis has discussed the interaction between FSM eligibility and WIMD. This study

Table 1
Characteristics of study population by FSM eligibility.

	FSM		Non-FSM		Difference (95% CI)
	N = 24,148	%	N = 134,983	%	
Gender					
Boy	12,175	50.4	68,704	50.9	
Girl	11,973	49.6	66,279	49.1	0.5(-0.2, 1.2)
Living area					
Urban	18,829	78.0	92,749	68.7	9.3(8.7, 9.8)
Rural	5319	22.0	42,234	31.3	
Number of adults in the household					
1	7062	29.2	17,639	13.1	16.2(15.6, 16.8)
2	9058	37.5	63,682	47.2	−9.7 (−10.3, −9.0)
3 and above	8028	33.2	53,662	39.8	−6.5 (−7.2, −5.9)
Number of other children in the household					
0	7079	29.3	57,438	42.6	−13.2 (−13.9, −12.6)
1	7557	31.3	50,774	37.6	−6.3 (−7.0, −5.7)
2	5036	20.9	18,878	14.0	6.9 (6.3, 7.4)
3 and above	4476	18.5	7893	5.8	12.7 (12.2, 13.2)
Living with someone who had alcohol problem					
No	21,499	89.0	129,874	96.2	
Yes	2649	11.0	5109	3.8	7.2 (6.8, 7.6)
Living with someone who had depression					
No	8865	36.7	81255	60.2	
Yes	15283	63.3	53728	39.8	23.5(22.8, 24.1)
Living with someone who had serious mental illness					
No	23,035	95.4	133,371	98.8	
Yes	1113	4.6	1612	1.2	3.4 (3.2, 3.7)
Exam year					
2009	2804	11.6	17,661	13.1	−1.5(-1.9, −1)
2010	2943	12.2	17,686	13.1	−0.9, (−1.4, −0.5)
2011	3125	12.9	17,247	12.8	.2(-0.3, 0.6)
2012	3039	12.6	16,779	12.4	.2(-0.3, 0.6)
2013	3428	14.2	17,511	13.0	1.2(0.8, 1.7)
2014	3110	12.9	16,836	12.5	0.4(-0.1, 0.9)
2015	2938	12.2	15,958	11.8	0.3(-0.1, 0.8)
2016	2761	11.4	15,305	11.3	0.1(-0.3, 0.5)
Special Education Need					
No	15,338	63.5	111,206	82.4	
Yes	8810	36.5	23,777	17.6	18.9 (18.3, 19.5)
Overall Welsh Index of Multiple Deprivation (WIMD)					
1 (Most deprived)	11,395	47.2	26,004	19.3	27.9(27.3, 28.6)
2	5891	24.4	26,724	19.8	4.6(4, 5.2)
3	3678	15.2	27,481	20.4	−5.1(-5.6, −4.6)
4	1865	7.7	24,686	18.3	−10.6(-11, −10.2)
5 (Least deprived)	1319	5.5	30,088	22.3	−16.8(17.2, −16.5)
Income WIMD					
1 (Most deprived)	11,439	47.4	25,539	18.9	28.5(27.8, 29.1)
2	6071	25.1	27,613	20.5	4.7(4.1, 5.3)
3	3599	14.9	27,105	20.1	−5.2(-5.7, −4.7)
4	2018	8.4	26,627	19.7	−11.4(-11.8, −11)
5 (Least deprived)	1021	4.2	28,099	20.8	−16.6(-16.9, −16.3)
Health WIMD					
1 (Most deprived)	10,173	42.1	26,465	19.6	22.5(21.9, 23.2)
2	6359	26.3	27,836	20.6	5.7 (5.1, 6.3)
3	3963	16.4	27,315	20.2	−3.8(-4.3, −3.3)
4	2281	9.4	25,830	19.1	−9.7(-10.1, −9.3)
5 (Least deprived)	1372	5.7	27,537	20.4	−14.7(-15.1, −14.4)
Access to service WIMD					
1 (Most deprived)	1834	7.6	23,492	17.4	−9.8(-10.2, −9.4)
2	3926	16.3	30,234	22.4	−6.1(-6.7, −5.6)
3	6206	25.7	28,194	20.9	4.8(4.2, 5.4)
4	6640	27.5	28,758	21.3	6.2(5.6, 6.8)
5 (Least deprived)	5542	23.0	24,305	18.0	4.9(4.4, 5.5)
Community safety WIMD					
1 (Most deprived)	9828	40.7	24,835	18.4	22.3(21.7, 23.0)
2	6293	26.1	27,291	20.2	5.8(5.3, 6.4)
3	4386	18.2	27,722	20.5	−2.4(-2.9, −1.8)
4	2429	10.1	28,324	21.0	−10.9(-11.4, −10.5)
5 (Least deprived)	1212	5.0	26,811	19.9	−14.8(15.2, 14.5)
Physical environment WIMD					
1 (Most deprived)	5501	22.8	26,282	19.5	3.3(2.7, 3.9)
2	4786	19.8	28,204	20.9	−1.1(-1.6, −0.5)
3	4866	20.2	28,320	21.0	−0.8(-1.4, −0.3)
4	4256	17.6	25,648	19.0	−1.4(-1.9, −0.9)
5 (Least deprived)	4739	19.6	26,529	19.7	0.0(-0.6, 0.5)
Housing WIMD					
1 (Most deprived)	6185	25.6	22,805	16.9	8.7(8.1, 9.3)
2	5422	22.5	25,205	18.7	3.8(3.2, 4.4)
3	5338	22.1	26,437	19.6	2.5(2, 3.1)
4	4756	19.7	27,487	20.4	−0.7(-1.2, −0.1)
5 (Least deprived)	2447	10.1	33,049	24.5	−14.4(-14.8, −13.9)

examined if a child's potential to leave poverty can be moderated by improvements in their in local built environment. This is measured by examining the association of the domains of WIMD (e.g. income, community safety, health, access to services, physical environment, housing) on a child's outcome in order to develop insight into the factors that best influence the child's trajectory. Logistic regression models were used to determine the association between local area deprivation measured by WIMD and achieving PLP amongst the children in Wales. The logistic regressor was augmented with stepwise bidirectional (forward and backward) search for optimal model selection (Burnham & Anderson, 2003). This method determines the best model with the minimum Akaike Information Criterion (AIC) and least significant features are excluded at each iteration step. The study has confirmed that there is no major concern around the high degree of correlation between predictor variables in the regression models by multicollinearity test (see Supplementary material collinearity test). Along with the explanatory variables, the stepwise logistic regression models have been adjusted for other covariates – such as exam year, gender, urban/rural classification of the living area, number of adults in the household, number of children in the household, living with someone who had an alcohol problem, living with someone who had depression, living with someone who had serious mental illness, child's special education need requirement – as these factors are also associated with the outcome variable. The odds ratio calculated with this adjustment has been reported throughout this work. The statistical significance of the explanatory variables and covariates have been interpreted by the p value less than 0.05. The data preparation including extraction, cleaning and linkage was performed in Structured Query Language (SQL) on an IBM DB2 platform and analyses were performed in the R statistical language version 3.3.2 (R Core Team, 2018).

Table 2
Breakdown of achieving PLP outcome variable.

	FSM		Non-FSM		Difference (95%CI)
	N = 24,148	%	N = 134,983	%	
PLP					
achieved	5311	22.0	74060	54.9	−32.9 (−33.5, −32.3)
not achieved	18837	78.0	60923	45.1	
KS4 not achieved:					
Achieved	6005	24.9	79083	58.6	−33.7 (−33.1,–34.3)
Not achieved	18143	75.1	55900	41.4	
Alcohol record					
No	22645	93.8	129441	95.9	−2.1 (−2.5, −1.8)
yes	1503	6.2	5542	4.1	
Substance misuse record					
No	23709	98.2	134130	99.4	−1.2 (−1.4, −1.0)
yes	439	1.8	853	0.6	
Any mental health condition					
No	21487	89.0	128172	95.0	−6.0 (−6.4, −5.6)
yes	2661	11.0	6811	5.0	

2.6Ethical approval

This study was approved by the SAIL Databank independent Information Governance Review Panel (IGRP) (project number 0916 – WECC Phase 4).

3. Results

Characteristics of the study population by family level poverty as assessed using FSM are presented in Table 1. Those receiving FSM were more likely (compared to non-FSM) to live in a single parent household (29.2% compared to 13.1%, respectively), live with 3 or more other children (18.5% compared to 5.8%, respectively) in the same household or to have special educational needs (36.5% compared to 17.6%). They were also more likely to live with a household member who had an alcohol problem (11% compared to 3.8%), depression (63.3% compared to 39.8%), or a serious mental illness (4.6% compared to 1.2%).

3.1. Outcomes for children on FSM

There were 22% FSM children who achieved PLP compared to 54.9% of non-FSM children (difference: 32.9% (95%CI: 32.3%, 33.5%)). Where children who did not achieve PLP this was mainly due to them not achieving KS4 (75.1% of children on FSM) and due to having a mental health condition (11% of FSM children) (see Table 2). The distribution of children for each component of the PLP across all WIMDs has been

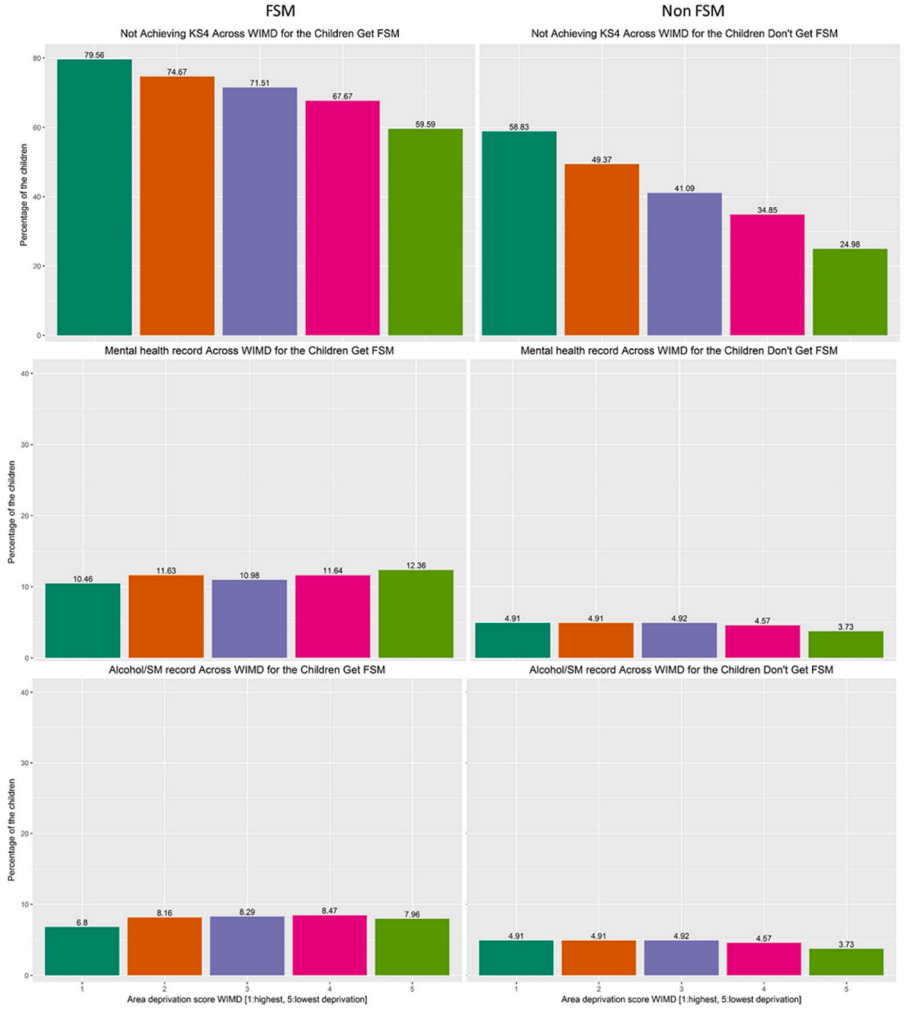


Fig. 2. The percentage of the children (FSM and non-FSM) for each component of the outcome variable PLP across all WIMDs.

Table 3
Logistic regression model of the association between overall WIMD and achieving PLP for the FSM and non-FSM children.

Variables	FSM children			non - FSM children		
	OR	Lower CI	Upper CI	OR	Lower CI	Upper CI
Overall WIMD						
1(Most deprived)	1.00			1.00		
2	1.27	1.17	1.38	1.40	1.35	1.45
3	1.46	1.32	1.60	1.88	1.81	1.95
4	1.76	1.56	1.98	2.34	2.25	2.44
5 (Least deprived)	2.20	1.93	2.51	3.47	3.34	3.61
Exam year						
2009	1.00			1.00		
2010	1.14	0.99	1.32	1.17	1.12	1.22
2011	1.29	1.12	1.49	1.23	1.18	1.29
2012	1.44	1.26	1.66	1.28	1.22	1.34
2013	1.67	1.46	1.91	1.44	1.37	1.50
2014	1.93	1.69	2.22	1.67	1.60	1.75
2015	2.31	2.01	2.65	1.93	1.84	2.03
2016	2.86	2.50	3.28	2.21	2.10	2.32
Gender						
Boys	–			1.00		
Girls	–			1.06	1.03	1.08
Living area						
Urban	1.00			1.00		
Rural	0.94	0.87	1.02	1.04	1.02	1.07
Number of adults in the household						
1	–			1.00		
2	–			1.48	1.43	1.53
3 and above	–			1.28	1.23	1.33
Number of children in the household						
0	1.00			1.00		
1	0.99	0.91	1.08	1.09	1.06	1.12
2	0.95	0.86	1.04	0.94	0.90	0.97
3 and above	0.88	0.80	0.98	0.79	0.75	0.83
Living with someone who had alcohol problem						
No	1.00			1.00		
Yes	0.77	0.68	0.86	0.62	0.58	0.66
Living with someone who had depression						
No	1.00			1.00		
Yes	0.88	0.82	0.94	0.69	0.67	0.71
Living with someone who had serious mental illness						
No	–			1.00		
Yes	–			0.89	0.80	1.00
Special Education Need						
No	1.00			1.00		
Yes	0.12	0.11	0.13	0.14	0.14	0.15

*Intercept for FSM model: 0.27(0.23, 0.30) and non-FSM model: 0.54(0.51, 0.57).

presented in Fig. 2.

3.2. Factors associated with achieving PLP for children who are on Free School Meals

Children who lived in a deprived household (based on FSM eligibility) but in the least deprived areas were significantly more likely to achieve PLP (adjusted odds ratio (aOR) 2.20 (1.93, 2.51)) compared to FSM children from the most deprived areas. Living in a household containing less than 3 children, and not living with someone with an alcohol problem or depression were also associated with achieving PLP for children living in high individual-level socio-economic deprivation (see Table 3). For FSM children gender, number of adult household members and living with someone who had serious mental illness were not as significantly associated with the outcome variable, as a result bidirectional model removed them in the iteration steps.

Supplementary work was conducted to investigate the association between WIMD and PLP components individually. It shows that poor children in least deprived areas are doing significantly better in education (aOR for achieving KS4 is 2.53 (2.23, 2.88)) than those living in most deprived areas. However, the trend is not similar for mental health, substance misuse and alcohol problems (see Supplementary Material Table 1).

3.3. Factors associated with achieving PLP for children who are not on Free School Meals

Like FSM children, non-FSM children who were living in the least deprived areas were also significantly more likely to achieve PLP (aOR 3.47 (3.34, 3.61)) compared to children living in deprived areas. Non-FSM girls were doing better than boys. The other most statistically significant factors that support these children to achieve were - not living in a single adult household, living with another child in the household and not living with someone with alcohol and mental health conditions.

The supplementary work (Supplementary Material Table 1) showed that non-FSM children living in least deprived areas were doing significantly better in all components of PLP than their peers from the most deprived areas, aOR for achieving KS4 is 3.91 (3.76, 4.06), aOR for not having mental health problems is 1.30 (1.21, 1.41), aOR for having substance misuse and alcohol problems is 1.21 (1.12, 1.32).

3.4. The impact of different aspects of area on achieving PLP for children on FSM

Children who were on FSM and living in deprived areas were significantly less likely to achieve PLP than children who were on FSM but living in less deprived areas (18.32% compared to 34.54%) (see Fig. 3). This figure suggests that despite household-level deprivation, children are able to achieve PLP if they are living in more affluent areas. The area components that made the most difference to children's achievement were higher community safety (1.95 times more likely to achieve for FSM children living in the safest areas compared to the least safe areas), higher relative income in the area (e.g. fewer people on benefits and more people in work, 1.61 times more likely to achieve if living in the highest income area compared to the lowest), and relatively higher access to services (1.26 times more likely to achieve if living in areas with high access to services compared to those with low access to services). After adjusting for WIMD domains children from urban areas were more likely to achieve compared to children from rural areas (see Table 4). Area characteristics that did not impact on achieving PLP included general health of people in the area or physical environment (e.g., pollution levels). Figs. 4 and 5 graphically depicts the significant indicators that were associated with achieving PLP for both FSM and non-FSM children.

3.5. Interaction between FSM and overall WIMD

Supplementary analysis describes the interaction between FSM eligibility and overall area-level deprivation as measured by WIMD (Supplementary Material Tables 2 and 3). The interaction model showed that FSM children in the least deprived areas were significantly more likely to achieve PLP than FSM children from the most deprived areas (aOR – 2.19 (1.92–2.50)). It also showed that FSM children living in the most deprived areas were less likely to achieve PLP than non-FSM children from similar areas.

4. Discussion

This study found that the area in which children grow up has an important impact on their developmental outcomes, especially at school, suggesting a neighbourhood effect on education irrespective or parental educational attainment (McDool, 2017). Previous research

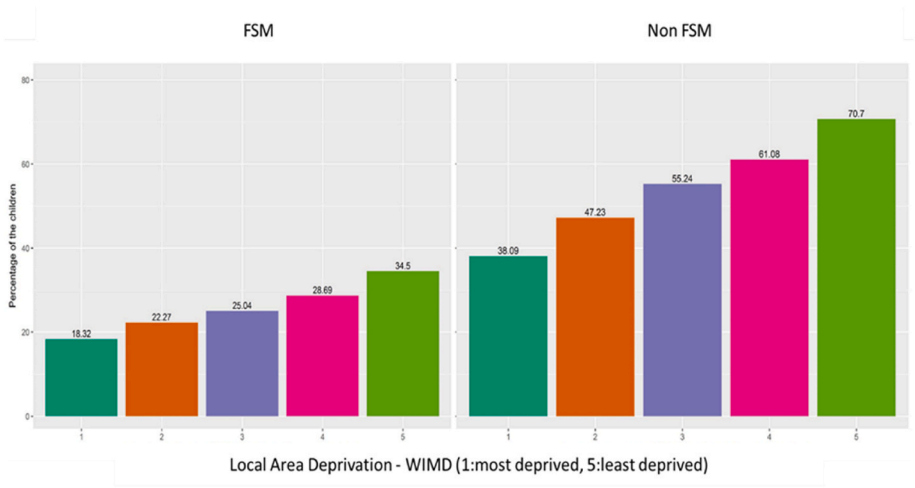


Fig. 3. The percentage of the children (FSM and non-FSM) who are achieving PLP across all area level deprivation scores.

suggested a relatively small association between neighbourhood effects and educational attainment and that family background is more of a factor (Gibbons, 2002). However, the findings from our study suggest that area level improvements have a positive impact on the outcome of the children, and it can moderate the effect of household level deprivation (Figs. 2 and 3). This trend is significant even after adjusting for other household-level factors (Tables 3 and 4). This study highlights specific aspects of neighbourhood characteristics e.g. community safety, area income and connectivity, which impact on children overcoming negative aspects of poverty. In terms of community safety, previous studies have showed that children are more able to undertake outdoor physical activity if they are living in a safer place and this directly contributes to their resilience (Flouri et al., 2014; Veitch et al., 2013). Other evidence indicates that concerns over community safety are a growing reason for dissatisfaction with green spaces (Welsh Government, 2018). Residents of deprived areas are more likely to report poorer safety in green spaces and visit them less frequently (Jones et al., 2009) with potential indirect consequences for physical development of children. Living in an area which feels unsafe due to high crime levels, has a detrimental effect on residents in general (Foster et al., 2013), hence living in an area with minimal crime risk becomes beneficial for child development. Evidence-based measures that improve area safely include neighbourhood watch, street lighting, CCTV, hotspots policing and alley gating (Crime Reduction Toolkit | College of Policing). In addition, this study also found that good access to services such as public transport, food shops, schools, leisure centres and health services are important aspects of the local area that helps children who are in poverty to achieve PLP in their life. There has been evidence that children living in an area with good access to services in day-to-day life has a positive influence on their overall development (Christian et al., 2015). An area with good public transport and good social connectivity is an advantageous environment for the children. This might explain why children in rural areas are less likely to achieve PLP than children in urban areas. The Income domain of WIMD reflects the proportion of the people who are living in the area who are claiming income-related benefits and qualitative evidence indicates that poverty also has an effect on children's experiences at school (Horgan, 2007). This study found that children who are in poverty (indicated by eligibility for FSM) do better when living in an area where fewer people are claiming benefits (e.g., less income-related deprivation in the area). This is also supported by a previous study which shows that if children in poverty have relocated to a less poor areas at an early stage, there is a decrease in the risk of adverse consequences in later life (Chetty et al., 2016). If the social norm is to be in employment this may make it also the 'norm' for children to remain in education or seek employment. The additional analysis conducted by the study found that the effect of local area on the child's educational attainment is clearer than its effect on child's mental health or alcohol or substance misuse, particularly for the children in household level deprivation (FSM children). This indicates that mental health and substance misuse might be more associated with individual level deprivation and factors within the family rather than local area and where are education is strongly associated with area level factors. This complex relationship needs further investigation.

This study brings together anonymously linked, routinely collected administrative datasets to build a nationally representative cohort of children and followed the study population longitudinally since birth till they complete KS4. This linked routine data framework facilitates the record linkage for the study population across health, education, and household level data. This is a major strength of the current study as this helps to overcome the limitations of selection and recall bias which are persistent in survey data. Also, data such as WIMD score, education record, FSM eligibility that were used to build the models in the study are available to government and policy-making bodies, hence these models can be exploited for developing intervention plans. However, the limitation of the study can be explained as this study uses person-level data to identify possible impact of non-income-based factors on child development and education outcomes. Aside from proxies for child poverty, such as FSM eligibility, the results indicate the effects of community safety, higher relative income, and access to services in an area on children's ability to achieve PLP. In doing so this study utilises small-area level measures from WIMD which are linked to ONS census geographies. ONS census geographies are designed to maintain best practice is disclosure controls for UK census data and therefore necessarily mask household-level variations in WIMD characteristics. This will introduce an ecological inference fallacy where aggregated data were used as a basis for individuals to make an inference (Hsieh, 2016). Aggregated-level data may not necessarily always be a true reflection of an individual; hence this can be a limitation of the study. However, the findings highlight the impact of broader area-related factors on child development in conjunction with a family level deprivation measure (FSM). Additionally, there is a need for multilevel modelling at various levels such as – LSOA, household and school, which would help to investigate the association of various granular area-level factors on a child's PLP profile. More than 20% of eligible children did not have a full GP record between age 11–16, so were excluded from the study. Also, those who did not have a continuous record in SAIL (WDSD dataset) from age 6 months were excluded. These children may have different characteristics of those included in the study and we cannot extrapolate to children who may have moved in or out of Wales in their early life, and the impact this has on achieving PLP, this might introduce a

Table 4
Logistic regression model of the association between WIMD components and achieving PLP for the FSM and non-FSM children.

Variables	FSM children			non-FSM children		
	OR	Lower CI	Upper CI	OR	Lower CI	Upper CI
Income WIMD						
1(Most deprived)	1.00			1.00		
2	1.00	0.91	1.11	1.22	1.17	1.27
3	1.19	1.04	1.36	1.39	1.32	1.46
4	1.39	1.16	1.67	1.70	1.60	1.81
5 (Least deprived)	1.61	1.26	2.05	2.14	1.99	2.31
Health WIMD						
1(Most deprived)	1.00			1.00		
2	1.02	0.94	1.12	1.06	1.02	1.10
3	1.11	0.99	1.25	1.12	1.07	1.17
4	0.96	0.82	1.12	1.06	1.01	1.12
5 (Least deprived)	0.89	0.73	1.09	1.12	1.05	1.18
Access to service WIMD						
1 (Most deprived)	1.00			1.00		
2	0.97	0.83	1.13	0.92	0.88	0.96
3	1.05	0.90	1.22	0.94	0.90	0.98
4	1.03	0.88	1.20	0.97	0.92	1.01
5 (Least deprived)	1.26	1.07	1.48	1.09	1.03	1.15
Community safety WIMD						
1 (Most deprived)	1.00			1.00		
2	1.16	1.05	1.27	1.11	1.07	1.16
3	1.37	1.22	1.54	1.24	1.18	1.30
4	1.47	1.25	1.72	1.38	1.30	1.46
5 (Least deprived)	1.95	1.57	2.42	1.69	1.58	1.81
Physical environment WIMD						
1 (Most deprived)	–			1.00		
2	–			1.02	0.98	1.06
3	–			0.97	0.93	1.00
4	–			0.98	0.94	1.02
5 (Least deprived)	–			0.98	0.95	1.02
Housing WIMD						
1 (Most deprived)	–			1.00		
2	–			1.04	1.00	1.08
3	–			1.06	1.02	1.11
4	–			1.10	1.06	1.15
5 (Least deprived)	–			1.17	1.11	1.23
Exam Year						
2009	1.00			1.00		
2010	1.13	0.98	1.31	1.17	1.12	1.23
2011	1.29	1.12	1.48	1.24	1.18	1.30
2012	1.44	1.25	1.65	1.29	1.23	1.35
2013	1.66	1.45	1.90	1.44	1.38	1.51
2014	1.94	1.69	2.23	1.68	1.60	1.76
2015	2.31	2.01	2.65	1.94	1.85	2.03
2016	2.83	2.47	3.25	2.21	2.11	2.32
Gender						
Boys	–			1.00		
Girls	–			1.06	1.03	1.08
Living area						
Urban	1.00			1.00		
Rural	0.88	0.81	0.96	0.94	0.91	0.97
Number of adults in the household						
1	–			1.00		
2	–			1.47	1.41	1.52
3 and above	–			1.27	1.22	1.31
Number of children in the household						
0	1.00			1.00		
1	1.00	0.92	1.08	1.09	1.06	1.12
2	0.96	0.87	1.05	0.94	0.91	0.97
3 and above	0.89	0.81	0.99	0.79	0.75	0.83
Living with someone who had alcohol problem						
No	1.00			1.00		
Yes	0.77	0.69	0.86	0.63	0.59	0.67
Living with someone who had depression						
No	1.00			1.00		
Yes	0.88	0.82	0.94	0.70	0.68	0.71
Living with someone who had serious mental illness						
No	–			1.00		
Yes	–			0.88	0.79	0.99
Special Education Need						
No	1.00			1.00		
Yes	0.12	0.11	0.13	0.14	0.14	0.15

*Intercept for FSM model: 0.24(0.20–0.29) and non-FSM model: 0.51 (0.47–0.55).

selection bias in the study. In some cases, there is the possibility that people may select areas, such as those performing well academically may move to be closer to good schools/libraries and that it is the people who chose the area rather than the influence of the area that impacts on education outcomes.

In summary, children who grow up in poverty but have achieved PLP as defined by; achieving qualifications at age 16 and do not have a mental health diagnosis or substance misuse (including alcohol) problems, are those who live in an area with good community safety, have good public transport and access to services and live in an area where people are employed rather than on benefits. The findings of the study are indicative of the fact that intervention in various aspects of a local area such as improving safety, connectivity and more people at work might help local children to achieve PLP in terms of education, mental health and reducing risk-taking behaviours (alcohol/drug use).

Ethical approval

This study was approved by the SAIL Databank independent Information Governance Review Panel (IGRP) (project number 0916 – WECC Phase 4).

Funding

This research has been carried out as part of the ADR Wales programme of work. The ADR Wales programme of work is aligned to the priority themes as identified in the Welsh Government's national strategy: Prosperity for All. ADR Wales brings together data science experts at Swansea University Medical School, staff from the Wales Institute of Social and Economic Research, Data and Methods (WISERD) at Cardiff University and specialist teams within the Welsh Government to develop new evidence which supports Prosperity for All by using the SAIL Databank at Swansea University, to link and analyse anonymised data. ADR Wales is part of the Economic and Social Research Council (part of UK Research and Innovation) funded ADR UK (grant ES/S007393/1).

This work was also supported by the National Centre for Population Health and Well-Being Research (NCPHWR) which is funded by Health and Care Research Wales. This work was supported by Health Data Research UK which receives its funding from HDR UK Ltd (NIWA1)

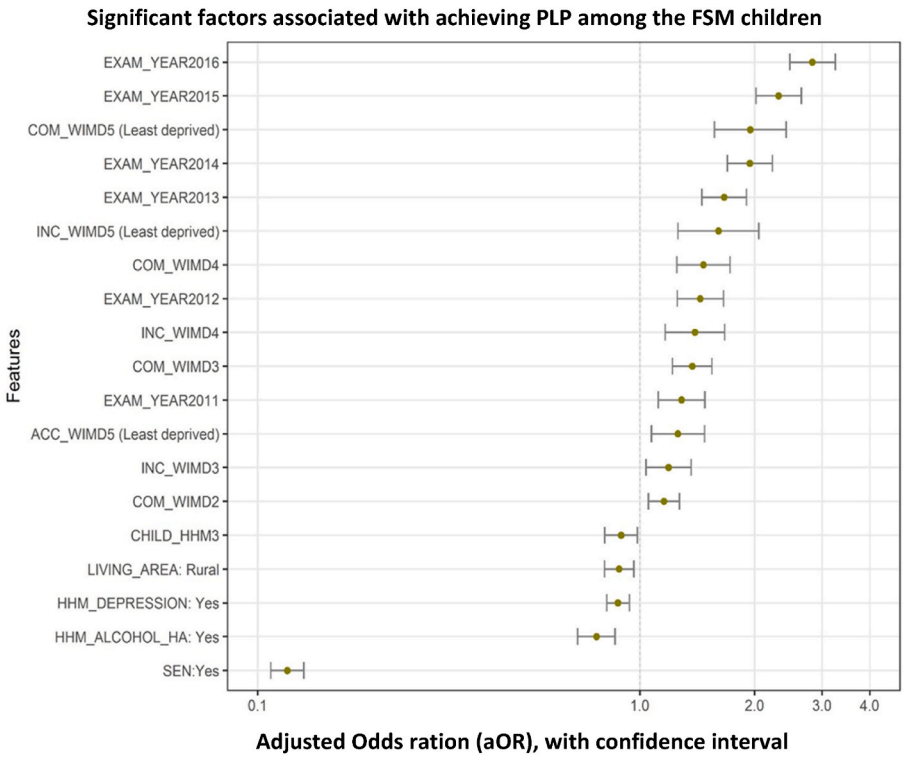


Fig. 4. Significant factors associated with achieving PLP among the FSM children. Note: EXAM_YAER = Exam year (between 2009 and 2016), COM_WIMD = Community safety WIMD, INC_WIMD = Income WIMD, ACC_WIMD = Access to service WIMD, CHILD_HHM = Number of children in the household, LIVING_AREA = Living area, HHM_DEPRESSION = Living with someone who had depression, HHM_ALCOHOL = Living with someone who had depression, SEN = Special Education Need.

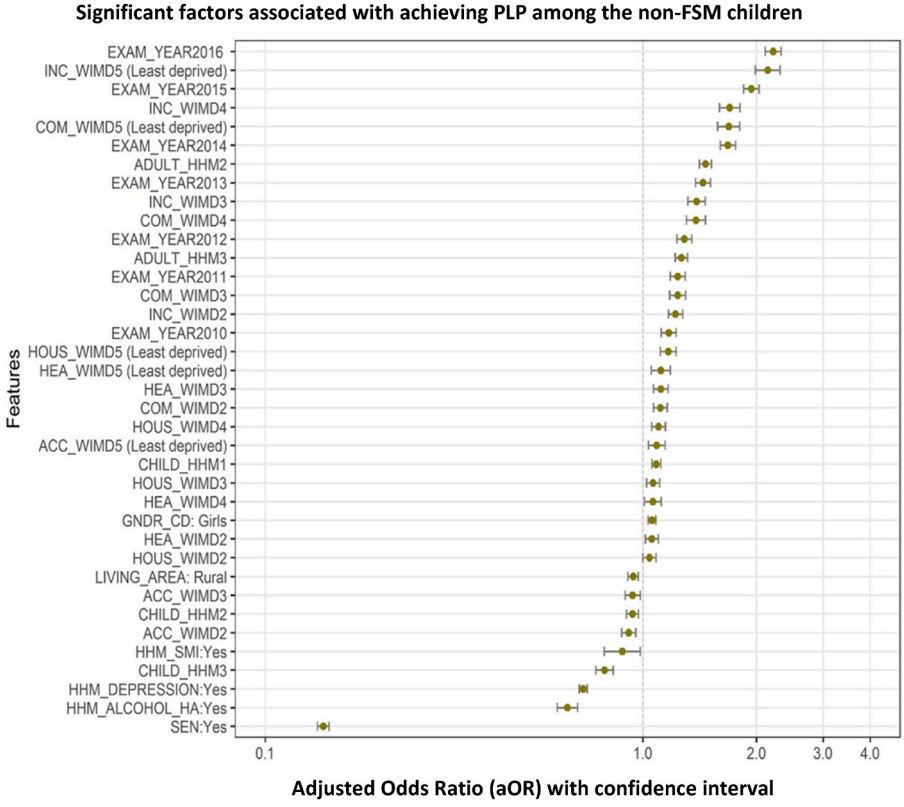


Fig. 5. Significant factors associated with achieving PLP among the non-FSM children. Note: EXAM_YAER = Exam year (between 2009 and 2016), INC_WIMD = Income WIMD, COM_WIMD = Community safety WIMD, ADULT_HHM = Number of adults in the household, HOUS_WIMD = Housing WIMD, HEA_WIMD = Health WIMD, ACC_WIMD = Access to service WIMD, CHILD_HHM = Number of children in the household, GNDR_CD = Gender, LIVING_AREA = Living area, HHM_SMI = Living with someone who had serious mental illness, HHM_DEPRESSION = Living with someone who had depression, HHM_ALCOHOL = Living with someone who had depression, SEN = Special Education Need.

funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Welcome Trust.

This work uses data provided by patients and collected by the NHS as part of their care and support. This study used anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. We

would like to acknowledge all the data providers who enable SAIL to make anonymised data available for research.

Authors' contributions

All authors contributed to the study conception and design. Data collection, preparation, and analysis were performed by Amrita Bandyopadhyay. The first draft of the manuscript was written by Amrita Bandyopadhyay. Tony Whiffen, Richard Fry and Sinead Brophy reviewed and edited the drafts. All authors read and approved the final manuscript. Conceptualization: Sinead Brophy and Amrita Bandyopadhyay; Methodology: Amrita Bandyopadhyay and Sinead Brophy; Formal analysis and investigation: Amrita Bandyopadhyay Writing - original draft preparation: Amrita Bandyopadhyay; Writing - review and editing: Tony Whiffen, Richard Fry and Sinead Brophy, Supervision: Sinead Brophy.

Participant consent

The study did not require participant consent as it utilises the anonymised data.

Patient and public involvement statement

No patient involved.

The original protocol

Not applicable.

STROBE checklist

STROBE checklist has been added as a Supplementary file (Supplementary material STROBE checklist).

Data sharing statement

The data have been archived in the Secure Anonymised Information Linkage Databank (https://saildatabank.com/0029).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Prof Sinead Brophy reports financial support was provided by Administrative Data Research Wales.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssmph.2023.101370.

References

Bandyopadhyay, A., Brophy, S., Akbari, A., Demmler, J., Kennedy, J., Paranjothy, S., Lyons, R., & Moore, S. (2022). Health and household environment factors linked with early alcohol use in adolescence: A record-linked, data-driven, longitudinal cohort study. *International Journal of Population Data Science, 7*(1), Article 1. https://doi.org/10.23889/ijpds.v7i1.1717

Beatty, C., & Fothergill, S. (2016). *The uneven impact of welfare reform: The financial losses to places and people*. Sheffield Hallam University.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer Science & Business Media.

Cecil-Karb, R., & Grogan-Kaylor, A. (2009). Childhood body mass index in community context: Neighborhood safety, television viewing, and growth trajectories of BMI. *Health & Social Work, 34*(3), 169–177. https://doi.org/10.1093/hsw/34.3.169

Chetty, R., Hendren, N., & Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *The American Economic Review; Nashville, 106*(4), 855–902. https://doi.org/10.1257/aer.20150572

Christian, H., Zubrick, S. R., Foster, S., Giles-Corti, B., Bull, F., Wood, L., Knuiman, M., Brinkman, S., Houghton, S., & Boruff, B. (2015). The influence of the neighborhood physical environment on early child health and development: A review and call for research. *Health & Place, 33*, 25–36. https://doi.org/10.1016/j.healthplace.2015.01.005

Cook, S. M. C., & Cameron, S. T. (2017). Social issues of teenage pregnancy. *Obstetrics, Gynaecology and Reproductive Medicine, 27*(11), 327–332. https://doi.org/10.1016/j.ogrm.2017.08.005

Cribb, J., Norris Keiller, A., & Waters, T. (2018). *Living standards, poverty and inequality in the UK: 2018 (issue R145)*. IFS Report.

Featherstone, B., Morris, K., Daniel, B., Bywaters, P., Brady, G., Bunting, L., Mason, W., & Mirza, N. (2019). Poverty, inequality, child abuse and neglect: Changing the conversation across the UK in child protection? *Children and Youth Services Review, 97*, 127–133. https://doi.org/10.1016/j.childyouth.2017.06.009

Flouri, E., Midouhas, E., & Joshi, H. (2014). The role of urban neighbourhood green space in children's emotional and behavioural resilience. *Journal of Environmental Psychology, 40*, 179–186. https://doi.org/10.1016/j.jenvp.2014.06.007

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), 157. https://doi.org/10.1186/1472-6963-9-157

Foster, S., Wood, L., Christian, H., Knuiman, M., & Giles-Corti, B. (2013). Planning safer suburbs: Do changes in the built environment influence residents' perceptions of crime risk? *Social Science & Medicine, 97*, 87–94. https://doi.org/10.1016/j.socscimed.2013.08.010

Galster, G., Marcotte, D. E., Mandell, M., Wolman, H., & Augustine, N. (2007). The influence of neighborhood poverty during childhood on fertility, education, and earnings outcomes. *Housing Studies, 22*(5), 723–751. https://doi.org/10.1080/02673030701474669

Gibbons, S., … (2002). *Neighbourhood effects on educational achievement (Issue 18)*. Centre for the Economics of Education, London School of Economics and.

Hingson, R. W., Heeren, T., & Winter, M. R. (2006). Age at drinking onset and alcohol dependence: Age at onset, duration, and severity. *Archives of Pediatrics and Adolescent Medicine, 160*(7), 739–746. https://doi.org/10.1001/archpedi.160.7.739

Horgan, G. (2007). *The impact of poverty on young children's experience of school* (Citeseer).

Hsieh, J. J. (2016). Ecological fallacy | epidemiology. Encyclopedia Britannica https://www.britannica.com/science/ecological-fallacy.

Johnson, R. D., Griffiths, L. J., Hollinghurst, J. P., Akbari, A., Lee, A., Thompson, D. A., Lyons, R. A., & Fry, R. (2021). Deriving household composition using population-scale electronic health record data—a reproducible methodology. *PLoS One, 16*(3), Article e0248195. https://doi.org/10.1371/journal.pone.0248195

Jones, A., Hillsdon, M., & Coombes, E. (2009). Greenspace access, use, and physical activity: Understanding the effects of area deprivation. *Preventive Medicine, 49*(6), 500–505. https://doi.org/10.1016/j.ypmed.2009.10.012

Joyce, R. (2014). *Child poverty in britain: Recent trends and future prospects* (working paper W15/07). IFS Working Papers https://doi.org/10.1920/wp.ifs.2015.1507.

Lambie-Mumford, H., & Green, M. A. (2017). Austerity, welfare reform and the rising use of food banks by children in England and Wales. *Area, 49*(3), 273–279. https://doi.org/10.1111/area.12233

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

Marco, A. D., & Vernon-Feagans, L. (2013). Rural neighborhood context, child care quality, and relationship to early language development. *Early Education & Development, 24*(6), 792–812. https://doi.org/10.1080/10409289.2013.736036

McDool, E. M. (2017). Neighbourhood effects on educational attainment: Does family background influence the relationship? *Sheffield Economics Research Papers*, Article 2017002. SERPS), 201700.

Minh, A., Muhajarine, N., Janus, M., Brownell, M., & Guhn, M. (2017). A review of neighborhood effects and early child development: How, where, and for whom, do neighborhoods matter? *Health & Place, 46*, 155–174. https://doi.org/10.1016/j.healthplace.2017.04.012

Observatory, P. H. W. (n.d.). *Public Health Wales Observatory—Overview—Child profile*. Public Health Wales Observatory. Retrieved August 5, 2020, from http://www.publichealthwalesobservatory.wales.nhs.uk/child-profile-overview.

Penman-Aguilar, A., Carter, M., Snead, M. C., & Kourtis, A. P. (2013). Socioeconomic disadvantage as a social determinant of teen childbearing in the U.S. *Public Health Reports, 128*(Suppl 1), 5–22. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3562742/.

R Core Team. (2018). *R: A language and environment for statistical computing*. https://www.r-project.org/.

Sattler, K., & Gershoff, E. (2019). Thresholds of resilience and within- and cross-domain academic achievement among children in poverty. *Early Childhood Research Quarterly, 46*, 87–96. https://doi.org/10.1016/j.ecresq.2018.04.003

Taylor, C. (2018). The reliability of free school meal eligibility as a measure of socio-economic disadvantage: Evidence from the millennium cohort study in Wales. *British Journal of Educational Studies, 66*(1), 29–51. https://doi.org/10.1080/00071005.2017.1330464

Taylor-Robinson, D., Lai, E. T. C., Wickham, S., Rose, T., Norman, P., Bambra, C., Whitehead, M., & Barr, B. (2019). Assessing the impact of rising child poverty on the unprecedented rise in infant mortality in England, 2000–2017: Time trend analysis. *BMJ Open, 9*(10), Article e029424. https://doi.org/10.1136/bmjopen-2019-029424

Loughborough University. (2019). *Child poverty growing fastest in the UK's most deprived areas*. Loughborough University. https://www.lboro.ac.uk/media-centre/press-releases/2019/may/child-poverty-growing-fasted-in-deprived-areas/.

Veitch, J., Hume, C., Salmon, J., Crawford, D., & Ball, K. (2013). What helps children to be more active and less sedentary? Perceptions of mothers living in disadvantaged neighbourhoods. *Child: Care, Health and Development, 39*(1), 94–102. https://doi.org/10.1111/j.1365-2214.2011.01321.x

Vizard, P., Obolenskaya, P., & Burchardt, T. (2019). Child poverty amongst young carers in the UK: Prevalence and trends in the wake of the financial crisis, economic downturn and onset of austerity. *Child Indicators Research, 12*(5), 1831–1854. https://doi.org/10.1007/s12187-018-9608-6

Welsh Government. (2011). *Welsh index of Multiple deprivation (full index update with ranks): 2011*. GOV.WALES. https://gov.wales/welsh-index-multiple-deprivation-full-index-update-ranks-2011.

Welsh Government. (2018). *National Survey for Wales 2017-18 play. Satistics for Wales*. https://duckduckgo.com/?t=ffab&q=National+Survey+for+Wales+2017-18Play&ia=web.

Welsh Government. (2019a). *Welsh index of Multiple deprivation (full index update with ranks): 2019*. GOV.WALES. https://gov.wales/welsh-index-multiple-deprivation-full-index-update-ranks-2019.

Welsh Government. (2019b). *Most children in poverty living in working households – new report*. GOV.WALES. https://gov.wales/most-children-poverty-living-working-households-new-report.

Wickham, S., Anwar, E., Barr, B., Law, C., & Taylor-Robinson, D. (2016). Poverty and child health in the UK: Using evidence for action. *Archives of Disease in Childhood, 101*(8), 759–766. https://doi.org/10.1136/archdischild-2014-306746

Wood, D. (2003). Effect of child and family poverty on child health in the United States. *Pediatrics, 112*(Supplement 3), 707–711. https://pediatrics.aappublications.org/content/112/Supplement_3/707.

My input

In this paper, I contributed to the development of a comprehensive research plan and conducted extensive data linkage, harmonisation and cleaning across all routine datasets for analysis within the SAIL Databank platform. I constructed a stepwise LR model to evaluate the relationships between local area characteristics and child outcomes. Furthermore, I authored and published this journal article as both the first author and corresponding author, ensuring that the findings and insights derived from this research reach a wider audience and inform future interventions in the field of early-life vulnerability profiling.

Impact

- This article has been published in *The SSM – Population Health* in 2023.
- The paper has been cited in seven other published works (google scholar).
- This paper constitutes a significant contribution to the early years of Administrative Data Research Wales (ADR Wales). The research question examined in this paper has been meticulously aligned with the strategic objectives of both ADR Wales and Welsh Government policies concerning early years. Furthermore, the findings have been communicated to both parties in the form of a comprehensive report.

Conclusion

The findings of this study highlight the significant influence of local area deprivation on children's life chances. Economic disadvantage at both the family and neighbourhood levels is associated with poorer health, lower educational attainment and higher risks of adverse social outcomes. However, the study also identifies factors that promote resilience among children in deprived areas, including strong community networks and access to early intervention services. These insights can help policymakers design more effective place-based interventions to support children living in poverty. Beyond socio-economic and environmental influences, exposure to domestic abuse (DA) is another key determinant of childhood vulnerability. The next chapter examines the intersection of DA and child outcomes, utilising linked data to understand risk and resilience patterns.

Chapter 5: Insights from linking police domestic abuse data and health data in South Wales, UK: a linked routine data analysis using decision tree classification

Critical summary

Background

DA is a major public health concern, as it causes long-term damage to victims and their families (75,76). Exposure to DA often leads to physical and psychological impairments (77) in children and is considered as a form of child maltreatment (76). Research has indicated that two third of female victims are revictimised, leading to increased utilisation of healthcare services, such as emergency department (ED) attendance (78). However, the vulnerability of these situations is often not disclosed to the ED (79). As a result, the circumstances remain unknown and the severity of the vulnerability escalates. The objective of this study is to gain comprehensive insights using an integrated data framework (WECC Phase 4) that enables data linkage between healthcare datasets and Public Protection Notification (PPN) police data. PPN data contains information on victims of DA through the DASH (Domestic Abuse, Stalking and Harassment) questionnaire (80). The DASH questionnaire, frequently used by the police and other agencies across the UK, is a tool to assess the risk of harm associated with victims of domestic abuse, stalking and harassment. The questions asked through the DASH questionnaire aim to capture the most important aspects of the situation and evaluate the severity of the condition.

Utilisation of administrative data

This study received the PPN DASH data for the population of South Wales from 2015 to 2020 from South Wales Police, one of the four police departments in Wales. South Wales Police agreed to provide their data to conduct this data linkage research project to identify the risk patterns of victims of DA and improve the ascertainment of individuals at high risk of DA without disclosure. This data was obtained with the aim of conducting an anonymised data linkage between victims of DA and their healthcare information including primary care from WLGP, hospital records from PEDW, and A&E attendance records from EDDS as well as administrative records (demographic data from WDS and death records from ONS). This data linkage between PPN DASH and health and administrative data has happened for the first time in Wales, enhancing the novelty of my research work.

In this study, EDDS, PEDW and ONS death datasets were utilised to derive the outcome variable. The primary outcome variable focused on any adverse outcomes, including ED attendance, emergency hospital admissions, or death due to any cause within 12 months

after the reported DA incident. This study also incorporated detailed information related to DA incidents (such as perpetration behaviour, injury details, victim's response) reported in the PPN DASH, which was utilised in the statistical analysis to investigate the profiles of victims who are more likely to experience adverse outcomes post-DA. The extensive use of administrative data allows for a comprehensive investigation of routine data to measure the impact of DA on victims, which is information that is otherwise difficult to obtain from other available systems.

Application of data science methods

A survival analysis was first conducted using Kaplan-Meier survival methods to evaluate the time to adverse outcomes for the victims. A multivariable Cox proportional hazards model was then developed to estimate the hazard ratios of the main risk factors contributing to these adverse outcomes. Following this, DT models were created to identify specific groups at risk of experiencing a DA incident and its subsequent outcomes. The DT models highlighted nodes related to the relevant risk profiles of the at-risk population through a graphical approach. This methodological approach is significant as it applies advanced statistical techniques to uncover the dynamics of vulnerability and its consequences. By applying survival analysis and DT modelling, the study builds critical risk profiles using visual representations of the data, making it easier to understand the complex relationships between exposure and outcome variables.

Early-life vulnerability profiling

The findings of the study propose a novel approach to identifying the most vulnerable at-risk populations by using routine data obtained from various services. In this study, the at-risk population is defined as those who are more likely to encounter a severe outcome after a domestic abuse (DA) incident, which includes emergency hospital admission, emergency department attendance and death within 11 months of the DA incident. The study established risk factors indicating vulnerability, suggesting that when a perpetrator has a history of violence, the victims are more likely to be at risk of future harm. Along with adults, this study investigated post-DA healthcare adverse outcomes among children (1.2% in the 0 to 9 age group and 10.7% among those aged 10 to 19); however, DA records were more prevalent among those aged 20 to 29 (26.5%). Findings revealed that, when comparing outcomes with the base age group of 20–29 years, children aged 10–19 years had a significantly higher risk of experiencing an adverse outcome (results are discussed in the main paper). This indicates that children are particularly susceptible to adverse outcomes and are likely to seek contact with the healthcare system after a DA incident, as their developmental needs may make them more vulnerable. Additionally, pregnant individuals face higher risks related to DA, which can affect both their health and that of their unborn child (this has been mentioned in chapter 2). Since this study shows that individuals involved in household abuse cases frequently have prior contact with various service providers, including general practitioners (GPs) and emergency

departments, it is valuable for identifying children and pregnant individuals who are linked to early-year vulnerability such as emotional developmental risk, physical health risk and lifelong impacts of DA exposure.

Published journal paper

Insights from linking police domestic abuse data and health data in South Wales, UK: a linked routine data analysis using decision tree classification

Natasha Kennedy, Tint Lwin Win, Amrita Bandyopadhyay, Jonathan Kennedy, Benjamin Rowe, Cynthia McNerney, Julie Evans, Karen Hughes, Mark A Bellis, Angela Jones, Karen Harrington, Simon Moore, Sinead Brophy

Summary

Background Exposure to domestic abuse can lead to long-term negative impacts on the victim's physical and psychological wellbeing. The 1998 Crime and Disorder Act requires agencies to collaborate on crime reduction strategies, including data sharing. Although data sharing is feasible for individuals, rarely are whole-agency data linked. This study aimed to examine the knowledge obtained by integrating information from police and health-care datasets through data linkage and analyse associated risk factor clusters.

Methods This retrospective cohort study analyses data from residents of South Wales who were victims of domestic abuse resulting in a Public Protection Notification (PPN) submission between Aug 12, 2015 and March 31, 2020. The study links these data with the victims' health records, collated within the Secure Anonymised Information Linkage databank, to examine factors associated with the outcome of an Emergency Department attendance, emergency hospital admission, or death within 12 months of the PPN submission. To assess the time to outcome for domestic abuse victims after the index PPN submission, we used Kaplan-Meier survival analysis. We used multivariable Cox regression models to identify which factors contributed the highest risk of experiencing an outcome after the index PPN submission. Finally, we created decision trees to describe specific groups of individuals who are at risk of experiencing a domestic abuse incident and subsequent outcome.

Findings After excluding individuals with multiple PPN records, duplicates, and records with a poor matching score or missing fields, the resulting clean dataset consisted of 8709 domestic abuse victims, of whom 6257 (71·8%) were female. Within a year of a domestic abuse incident, 3650 (41·9%) individuals had an outcome. Factors associated with experiencing an outcome within 12 months of the PPN included younger victim age (hazard ratio 1·183 [95% CI 1·053–1·329], p=0·0048), further PPN submissions after the initial referral (1·383 [1·295–1·476]; p<0·0001), injury at the scene (1·484 [1·368–1·609]; p<0·0001), assessed high risk (1·600 [1·444–1·773]; p<0·0001), referral to other agencies (1·518 [1·358–1·697]; p<0·0001), history of violence (1·229 [1·134–1·333]; p<0·0001), attempted strangulation (1·311 [1·148–1·497]; p<0·0001), and pregnancy (1·372 [1·142–1·648]; p=0·0007). Health-care data before the index PPN established that previous Emergency Department and hospital admissions, smoking, smoking cessation advice, obstetric codes, and prescription of antidepressants and antibiotics were associated with having a future outcome following a domestic abuse incident.

Interpretation The results indicate that vulnerable individuals are detectable in multiple datasets before and after involvement of the police. Operationalising these findings could reduce police callouts and future Emergency Department or hospital admissions, and improve outcomes for those who are vulnerable. Strategies include querying previous Emergency Department and hospital admissions, giving a high-risk assessment for a pregnant victim, and facilitating data linkage to identify vulnerable individuals.

Funding National Institute for Health Research.

Lancet Public Health 2023; 8: e629–38

See **Comment** page e580

National Centre for Population Health and Wellbeing Research (N Kennedy PhD, A Bandyopadhyay MSc, J Kennedy EngD, K Harrington MSc, Prof S Brophy PhD), **Health Data Research UK** (Prof S Brophy), **Administrative Data Research Wales** (N Kennedy, C McNerney BBus, Prof S Brophy), **and SAIL Databank** (C McNerney), **Swansea University Medical School** (T L Win MSc), **Swansea, UK; Data Lab, National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Swansea, UK** (J Kennedy); **South Wales Police, South Wales Police Head Quarters Cowbridge Road, Bridgend, UK** (B Rowe BSc); **Public Health Wales, Cardiff, UK** (J Evans MSc, Prof K Hughes PhD, A Jones MPH); **WHO Collaborating Centre for Violence Prevention, Liverpool John Moores University, Liverpool, UK** (Prof M A Bellis DSc); **Security, Crime & Intelligence Innovation Institute and Violence Research Group, School of Dentistry, Cardiff University, Heath Park, Cardiff, UK** (Prof S Moore PhD)

Correspondence to:
Dr Natasha Kennedy, National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Swansea SA2 8PP, UK
[redacted]

Introduction

Attending police officers issue a Public Protection Notification (PPN) to document the vulnerabilities of individuals who are victims of domestic abuse.[1] Once issued, the documented information is forwarded to the Force Public Protection Unit for a comprehensive risk assessment and subsequent determination of necessary action.[1] Although individuals have a right to privacy, a PPN enables information specific to the risk of serious harm to be shared with partner agencies when multi-agency management provides the appropriate response to those risks. Indeed, the 1998 Crime and Disorder Act requires agencies to collaborate on crime reduction strategies, including data sharing. The Domestic Abuse, Stalking, and Harassment (DASH) questionnaire is a standardised risk assessment tool adopted by the police and other agencies to identify and evaluate the potential risk of harm to victims of domestic abuse, stalking, or

Research in context

Evidence before this study

PubMed and Web of Science were searched for studies published in any language between April 17, 2013 and April 17, 2023, that investigated health-care use by victims of domestic abuse. Using the terms ("domestic violence" or "domestic abuse" or "intimate partner violence") and ("emergency medical attendance" or "emergency department visits" or "ED attendance" or "hospitalisation") and ("police"), the search yielded 198 results. The results were further filtered to include papers with the full text available and limited to research involving a human cohort. We manually searched the reference lists of the resulting studies for appropriate papers and highlighted key authors for relevant further studies. Numerous studies report the prevalence of domestic abuse victims and the negative health consequences that are endured. Consequently, these adverse health outcomes result in increased interactions with health-care services. However, knowledge is not communicated between agencies on a national level, preventing the implementation of safeguarding measures. Only four studies have investigated the value of linking data from police and health-care records on domestic abuse, with studies set in the USA and Canada. The findings of these studies suggested several risk factors for adverse outcomes, including increased Emergency Department usage, subsequent injuries when the incident involved physical violence, and the perpetrator having a history of domestic abuse. Furthermore, the studies showed that victims of domestic abuse often experience mental health disorders.

Added value of this study

In this study, we describe the value of unifying sensitive data, such as police and health-care data. This present study shows that highly vulnerable individuals frequently interact with health-care services but remain unknown to the police until a critical incident, underscoring the value in establishing data linkages across different agencies. Moreover, the study provides a use case that illustrates results that can be derived from linking whole data systems to identify points of early intervention for individuals at risk of domestic abuse, thus enabling proactive, upstream efforts to protect families.

Implications of all the available evidence

The evidence presented in this study demonstrates the preventive opportunities for stakeholders across multiple sectors, which can be facilitated when agencies communicate and link data. Future research should examine whether early identification of these vulnerable groups by the police or the health-care sector could lead to improved outcomes, in addition to examining the outcomes stratified by demographic profiles. This research can guide the development of targeted interventions that might mitigate the escalation of domestic abuse and related health outcomes.

harassment.[1,2] The DASH questionnaire comprises a series of methodically constructed questions that capture vital aspects of the situation, such as the perpetrator's behaviour, the victim's response, and the level of threat to the victim's safety. The collected data are analysed to determine the risk of further harm, and the results are used to implement effective interventions to safeguard the victim.[2]

Domestic abuse—defined here to include domestic violence and intimate partner violence, as per the DASH questionnaire—encompasses a range of incidents characterised by coercive, controlling, and abusive behaviours directed towards another individual.[3] These behaviours manifest in various forms, such as coercive control and emotional, financial, physical, psychological, or sexual violence or threats.[3,4] A lifetime incidence of domestic abuse has been estimated to affect 27% of women globally,[5] with a former intimate partner or spouse as the predominant perpetrator.[4]

Studies have indicated that exposure to domestic abuse results in long-term negative impacts on the victim's physical and psychological wellbeing.[6–8] Research into the long-term consequences of repeated victimisation is minimal; however, about two-thirds of female domestic abuse victims are revictimised.[4] When a cycle of revictimisation is present, it exacerbates the consequences of domestic abuse, such as affecting new relationships, introducing negative health behaviours in the form of addictions, and developing mental health disorders, including post-traumatic stress disorder.[4,7–10] This cycle of revictimisation also increases the use of health-care services. Domestic abuse victims interact more frequently with Emergency Departments than people from the general population.[4,11,12]

Although domestic abuse victims attend Emergency Departments frequently, these visits rarely result in the disclosure of abuse to staff or reports to the police.[11,12] Research into the identification of domestic abuse by Emergency Department staff indicated that, of 259 visits in 1999–2001 in a semi-rural county in the USA, physicians were more likely to document the violence (83% documented by the physician) than triage nurses (62%) or treatment nurses (44%).[11] In the context of emergency care, it is expected that clinical personnel proactively identify and address vulnerabilities that extend beyond the initial presenting symptoms. In addition to patients' disclosure, clinicians routinely assess their medical history using available data. Consequently, clinicians can enact safeguarding measures, including referring patients to an Independent Domestic Violence Advocate, if they perceive that the patient is at risk, even in cases where disclosure is absent. Further studies have indicated that domestic abuse victims frequently present to the Emergency Department

for non-injury-related complaints. Additionally, they also have an increased likelihood of having accompanying medical records documenting mental health and substance abuse issues.[11,13–15] Consequently, the Emergency Department and hospital setting offers a crucial opportunity for intervention and prevention, prioritising the health and wellbeing of domestic abuse victims. However, it is necessary to establish a robust infrastructure to gather additional information to effectively address the victims' needs.

This study aimed to examine knowledge obtained by integrating information from police PPN records and health-care datasets through data linkage. The study was conducted by investigating associations with information from numerous datasets, and examining the clustering patterns of related factors, as well as their association with Emergency Department attendance, emergency hospital admission, or death within 12 months of the index PPN submission, thereby identifying opportunities to improve ascertainment of individuals at high risk of domestic abuse without disclosure.

Methods

Study design and setting

We conducted a retrospective cohort study to ascertain the risk factors associated with victims who will experience an outcome (ie, Emergency Department admission, emergency admission to hospital, or death) in the 12 months following a PPN submission for domestic abuse. The cohort was composed of residents in the South Wales Police Force Region who were domestic abuse victims resulting in a PPN submission between Aug 12, 2015 and March 31, 2020. The cohort was based on the PPN dataset that also comprises records formed from the DASH risk identification and assessment model.[2] Data of anonymised identified persons were linked on the individual level to health record data within the Secure Anonymised Information Linkage (SAIL) databank.[16–18] The SAIL databank is a data repository containing over 10 billion anonymised records, with a population coverage of 100% for hospital and general practitioner (GP) data for this South Wales dataset, thus enabling person-based data linkage across numerous datasets. Each individual is assigned an encrypted anonymised linking field; this field is used to link anonymised individuals across datasets, thus facilitating longitudinal analysis of the individual's progression through the different datasets.[17] The linked data includes the primary care Wales Longitudinal General Practice dataset to identify reasons for contact with health-care professionals in general practice; data collected by GPs are captured via Read Codes, version 2, which relate to diagnosis, medication, and process-of-care codes. Hospital inpatient and outpatient data are collated in the Patient Episode Database for Wales, which encompasses clinical information pertaining to patients' hospital admissions, diagnoses, operations, and discharges using the International Classification of Diseases, 10th revision (ICD-10) clinical classification system. The Emergency Department Dataset for Wales uses three-digit alphanumeric codes to capture data regarding activity and information from Emergency Department and Minor Injury Units. The Office of National Statistics mortality dataset held within the Annual District Death Extract dataset contains demographic data, place of death, and underlying cause of death as ICD-10 codes. The Welsh Demographic Service dataset was used to identify all patients registered with a GP practice and to flag when people move in and out of Wales.

For the **Emergency Department Dataset for Wales** see https://web.www.healthdatagateway.org/dataset/75c4dcb8-33bf-43f4-b2bb-db51b6621b2c

For the **Annual District Death Extract dataset** see https://web.www.healthdatagateway.org/dataset/15cf4241-abad-4dcc-95b0-8cd7c02be999

For the **Welsh Demographic Service dataset** see https://web.www.healthdatagateway.org/dataset/8a8a5e90-b0c6-4839-bcd2-c69e6e8dca6d

Variables

The outcomes for all analyses were Emergency Department attendance, emergency hospital admission, and death due to any cause within 12 months of the index PPN submission; where the time to event is reported, it is the time to the first event for cases where an individual had multiple events (ie, emergency admission followed by death). An Emergency Department admission includes Emergency Department attendance or emergency admission to hospital. For women, we ensured that the hospital admissions and Emergency Department attendances were non-obstetric. Emergency Department, hospital, and GP data were examined for up to 1 and 3 years before receiving the index PPN submission to highlight early risk factors of experiencing an outcome following a domestic abuse incident (appendix 1 p 1).

See Online for appendix 1

In terms of exposures and confounders, all variables recorded at the index PPN from the DASH questionnaire[2] were included: attempted strangulation, conflict over a child (ie, conflict regarding contact with the child), hurt other people, history of further violence, injury, multi-agency risk assessment conference (MARAC) referral, pregnant, and past pregnancy. Subsequent PPN visits following the initial index PPN were also included as an explanatory variable. Additionally, for the decision tree analysis, information from Emergency Department and GP (ie, diagnoses, medications, procedures, and referrals) and hospital admissions (ie, cause of admission and date) were included in the analysis up to 1 and 3 years before the index PPN. Specific codes included are in the code list in appendix 2; these were filtered to include codes that possessed a frequency of greater than or equal to 250 within the cohort. The majority of the variables from the dataset were presented in a binary form, with 1 representing the presence of a concept and 0 representing its absence. The age of individuals was presented in 10-year brackets to understand the difference between different age groups while maintaining a suitable population size for analysis.

For the **Wales Longitudinal General Practice dataset** see https://web.www.healthdatagateway.org/dataset/33fc3ffd-aa4c-4a16-a32f-0c900aaea3d2

For more on **Read Codes** see https://digital.nhs.uk/services/terminology-and-classifications/read-codes

Data access

The data used in this study are available in the SAIL databank[18] at Swansea University (Swansea, UK). All data

For the **Patient Episode Database for Wales** see https://web.www.healthdatagateway.org/dataset/4c33a5d2-164c-41d7-9797-dc2b008cc852

See Online for appendix 2

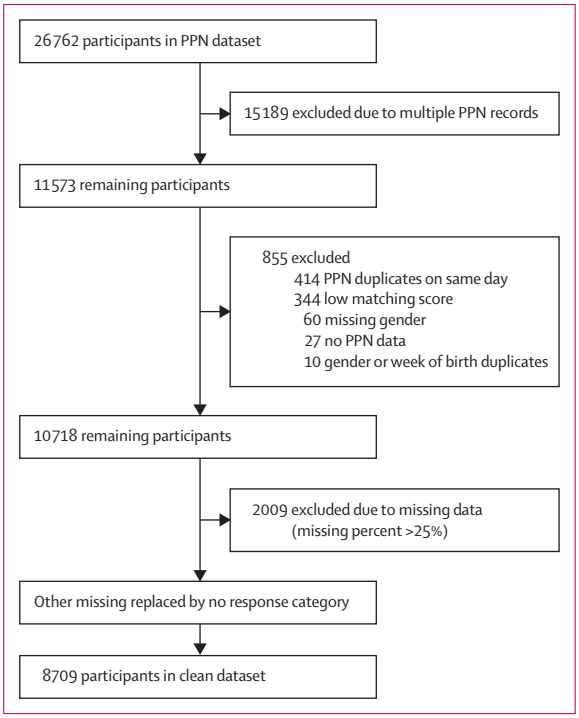


Figure 1: **Study profile**
PPN=Public Protection Notification.

held in the SAIL databank are anonymised; therefore, ethical approval is not mandatory in accordance with the Health Research Authority guidance and there is no legal requirement for explicit consent to participate under the Data Protection Act and UK General Data Protection Regulation. Furthermore, permission has been obtained from the relevant Caldicott Guardian or Data Protection Officer for all data contained in SAIL. In addition, proposals using SAIL data are subject to review by an Information Governance Review Panel (IGRP) to secure approval. The IGRP approval number for this study is 0916.

Statistical analysis

After the exclusion of multiple PPN records and records with missing fields or duplicates, missing values were visualised; individuals with 25% or more of missing data were removed. The missing values for the remaining individuals were replaced with a new category, NR (no response). Any categorical variables were altered to factors for analysis.

Descriptive statistics were generated to assess the rates of experiencing an outcome, stratified by both gender and age for the domestic abuse victims who were the subject of a PPN submission. We used Kaplan-Meier survival analysis to examine the time to outcome for the domestic abuse victim after the index PPN submission, censoring individuals who moved out of Wales. We used multivariable Cox proportional-hazard models, adjusted for confounders, to identify which factors contributed the highest risk of experiencing an outcome after the index PPN submission. The hazard ratios (HRs) were reported with 95% CIs and a significance level accepted at $p<0·05$. The reference groups were No for most factors; otherwise, the reference group was Male for gender, Standard Risk for the risk assessment factor, and the age group 20–29 years for age comparisons.

We created decision trees to identify specific groups of individuals who are at risk of experiencing a domestic abuse incident and subsequent outcome. The decision trees were not used to develop a predictive model, they were utilised in the context of descriptive epidemiology to investigate the clustering of risk factors within individuals (ie, everything represented by codes in the code list [appendix 2]). This approach facilitated the identification of the relevant nodes that could be used to classify an individual considering their risk factor clusters. The data handling and preparation were performed in SQL, using Eclipse 2020-03 (version 4.15). Final data preparation was performed in R Studio, version 4.1.3, whereas the decision trees were conducted in IBM SPSS, version 28.0.0.0.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

The PPN dataset contained 26762 records; of these, 15189 individuals had multiple PPN records. The exclusion of multiple PPN records resulted in a dataset comprising only the index PPN record, generating a cohort of 11573 domestic abuse victims. We also excluded records with a poor matching score (n=344; appendix 1 p 1), duplicates (n=424), missing gender (n=60), and an empty PPN record (n=27). Individuals with 25% or more of missing data were removed (n=2009), resulting in a 3% decrease of male victims, disproportionate to female victims, in the 20–39-year age range.

The resulting cohort comprised 8709 individuals (figure 1), of whom 6257 (71·8%) were female (table 1). The age groups of 20–29 years (30·0%) and 30–39 years (26·4%) contributed the highest proportion of domestic abuse victims; conversely, the age groups 0–9 (0·6%) and 80 years or older (0·7%) contributed the lowest proportion of domestic abuse victims.

3650 (41·9%) individuals had an outcome within a year of a domestic abuse incident; 2661 (72·9%) were female and 989 (27·1%) were male (table 1). 3544 individuals attended an Emergency Department, of whom 1085 individuals were then transferred to hospital. Overall, 1182 had an emergency hospital admission.

The age distribution of the outcome cohort indicates that more women than men had an outcome in the age range 20–39 years, which comprises the largest

proportion of domestic abuse victims. The total Emergency Department attendance rate was 572 (95% CI 554–591) per 1000 person-years and the hospitalisation rate was 185 (175–197) per 1000 person-years.

In comparison with the base age group of 20–29 years, those aged 10–19 years had an increased risk of experiencing an outcome (HR 1·183 [95% CI 1·053–1·329], p=0·0048; table 2; appendix 1 pp 1–4). Conversely, individuals aged 30–69 years had a decreased risk of experiencing an outcome (30–39 years: 0·882 [0·808–0·962], p=0·0045; 40–49 years: 0·801 [0·724–0·887], p<0·0001; 50–59 years: 0·772 [0·681–0·874], p<0·0001; 60–69 years: 0·728 [0·602–0·879], p=0·0010).

Cases involving attempted strangulation of the victim were associated with a higher risk of a future outcome (HR 1·311 [95% CI 1·148–1·497]; p<0·0001) than cases for which attempted strangulation was not present. Furthermore, victims had a higher risk of experiencing an outcome in the year proceeding a domestic abuse incident when the incident resulted in an injury (1·484 [1·368–1·609]; p<0·0001) than when it did not. Cases involving a pregnant household member had an increased risk of a future non-obstetric outcome (1·372 [1·142–1·648]; p=0·0007). Incidences where the perpetrator has hurt other people (1·218 [1·028–1·444]; p=0·023) or has a history of further violence (1·229 [1·134–1·333]; p<0·0001) resulted in an increased risk of the victim undergoing an outcome. Similarly, households that were subject to a MARAC referral (1·518 [1·358–1·697]; p<0·0001) or received multiple subsequent police visits after the index PPN (1·383 [1·295–1·476]; p<0·0001) have an increased risk of experiencing an outcome. Cases assessed as high risk, medium risk, or receiving no response from the responding police officer (high risk: 1·600 [1·444–1·773]; p<0·0001; medium risk: 1·117 [1·034–1·206]; p=0·0051; no response: 1·188 [1·075–1·312]; p=0·0007) were associated with a higher risk of experiencing an outcome than cases assessed as standard risk. Incidents involving conflict over a child had a lower risk of undergoing an outcome (0·856 [0·774–0·947]; p=0·0026) than those that did not. Furthermore, if a household member has had a child in the 18 months before the domestic abuse incident (past pregnancy), then the risk of an outcome is lowered (0·812 [0·722–0·913]; p=0·0005).

A decision tree combining knowledge gathered from GP and Emergency Department admissions up to a year before the domestic abuse incident, as well as information obtained by the police during the PPN submission, indicated that any Emergency Department admission before the domestic abuse incident is the most significant risk factor of experiencing an outcome (figure 2). Those with an Emergency Department admission 1 year before the event were further classified with the quantity (PPN count) and severity (MARAC referral) of their interactions with the police. Those who are known to the health-care system and go on to have further contact with the police were the most at risk of having an outcome after the domestic abuse incident.

	Male	Female	Total
Demographics			
Total cohort	n=2452	n=6257	n=8709
Age category, years			
0–9	29 (1·2%)	22 (0·4%)	51 (0·6%)
10–19	263 (10·7%)	521 (8·3%)	784 (9·0%)
20–29	649 (26·5%)	1962 (31·4%)	2611 (30·0%)
30–39	595 (24·3%)	1700 (27·2%)	2295 (26·4%)
40–49	423 (17·3%)	1085 (17·3%)	1508 (17·3%)
50–59	275 (11·2%)	617 (9·9%)	892 (10·2%)
60–69	137 (5·6%)	211 (3·4%)	348 (4·0%)
70–79	57 (2·3%)	104 (1·7%)	161 (1·8%)
≥80	24 (1·0%)	35 (0·6%)	59 (0·7%)
Outcomes			
Overall	n=989	n=2661	n=3650
Type of outcome			
Any admission	985 (99·6%)	2656 (99·8%)	3641 (99·8%)
Emergency Department attendance	960 (97·1%)	2584 (97·1%)	3544 (97·1%)
Emergency hospital admission	266 (26·9%)	916 (34·4%)	1182 (32·4%)
Death	23 (2·3%)	25 (0·9%)	48 (1·3%)
Outcomes by age category, years			
0–9	16 (1·6%)	9 (0·3%)	25 (0·7%)
10–19	110 (11·1%)	286 (10·7%)	396 (10·8%)
20–29	294 (29·7%)	876 (32·9%)	1170 (32·1%)
30–39	227 (23·0%)	701 (26·3%)	928 (25·4%)
40–49	167 (16·9%)	411 (15·4%)	578 (15·8%)
50–59	99 (10·0%)	228 (8·6%)	327 (9·0%)
60–69	43 (4·3%)	78 (2·9%)	121 (3·3%)
70–79	21 (2·1%)	55 (2·1%)	76 (2·1%)
≥80	12 (1·2%)	17 (0·6%)	29 (0·8%)

Table 1: **Descriptive statistics of the cohort split by gender**

Individuals with up to one Emergency Department admission and who had subsequent interactions with the police were further split by the prescription of CNS drugs; those who were prescribed these drugs were more likely to experience an outcome. These drugs were mostly prescribed for anxiety, depression, and sleep disorders. Individuals who were not prescribed CNS drugs were split on age; the extreme age groups, 10–19 years and 80 years and older, were classified as being at greater risk of undergoing an outcome after a domestic abuse incident than those in other age groups.

For individuals who had minimal interactions with the health-care and police systems, the question about history of further violence was important as it is asked when there is an injury at the scene. Therefore, in scenarios where this question was asked, the risk of a future outcome was higher than scenarios where this question was not asked and marked as NA (ie, not applicable).

	Hazard ratio (95% CI)	p value
Age group, years		
20–29	Ref	
0–9	1·400 (0·938–2·092)	0·10
10–19	1·183 (1·053–1·329)	0·0048
30–39	0·882 (0·808–0·962)	0·0045
40–49	0·801 (0·724–0·887)	<0·0001
50–59	0·772 (0·681–0·874)	<0·0001
60–69	0·728 (0·602–0·879)	0·0010
70–79	1·087 (0·860–1·374)	0·48
≥80	1·110 (0·767–1·607)	0·58
Attempted strangulation		
No	Ref	
No response	0·970 (0·602–1·562)	0·90
Yes	1·311 (1·148–1·497)	<0·0001
Conflict over a child		
No	Ref	
No response	1·060 (0·709–1·584)	0·78
Yes	0·856 (0·774–0·947)	0·0026
Hurt other people		
No	Ref	
No response	0·570 (0·285–1·140)	0·11
Yes	1·218 (1·028–1·444)	0·023
History of further violence		
No	Ref	
No response	0·905 (0·649–1·263)	0·56
Yes	1·229 (1·134–1·333)	<0·0001
Gender		
Male	Ref	
Female	1·071 (0·996–1·152)	0·066
Injury		
No	Ref	
No response	1·143 (0·571–2·288)	0·71
Yes	1·484 (1·368–1·609)	<0·0001
MARAC referral		
No	Ref	
No response	1·024 (0·956–1·097)	0·51
Yes	1·518 (1·358–1·697)	<0·0001
Multiple police visits (PPNs)		
No	Ref	
Yes	1·383 (1·295–1·476)	<0·0001
Past pregnancy		
No	Ref	
No response	0·784 (0·516–1·193)	0·26
Yes	0·812 (0·722–0·913)	0·0005
Pregnant		
No	Ref	
No response	0·929 (0·514–1·679)	0·81
Yes	1·372 (1·142–1·648)	0·0007
Risk assessment		
Standard	Ref	
No response	1·188 (1·075–1·312)	0·0007
Medium	1·117 (1·034–1·206)	0·0051
High	1·600 (1·444–1·773)	<0·0001

(Table 2 continues in next column)

(Continued from previous column)

MARAC=multi-agency risk assessment conference. PPN=Public Protection Notification.

Table 2: **Hazard ratios and CIs for each factor for predicting an emergency attendance following a PPN, adjusted for confounders**

The smoking status of the victim was indicated as an important factor for those for whom NA was the answer to history of further injury; the presence of smoking behaviour related to an increased risk of experiencing an outcome.

Incorporating additional information provided by the health-care system up to 3 years preceding the index PPN submission further indicated that multiple Emergency Department admissions before the domestic abuse incident was significantly associated with experiencing an outcome (figure 3). The individuals who were most at risk of undergoing an outcome after the domestic abuse incident were those who had more than seven Emergency Department admissions in the 3 years before and who received smoking cessation advice from their GP. Individuals who had more than three Emergency Department admissions but less than seven admissions, and have had more than two PPN submissions, are also at high risk; those who were classified as a cigarette smoker by their GP had a greater risk of experiencing an outcome than those who were not. Individuals with less than three Emergency Department admissions were split by future PPN count; overall, those with a PPN count of less than one were at the lowest risk of having an outcome compared with the entire cohort. However, those individuals who were not known to the health-care or police systems but had been prescribed CNS drugs by their GP and who were in the extreme age group categories of 0–29 years and 70–79 years were at risk of undergoing an outcome.

Discussion

This study shows how communication between separate services can be utilised to identify points of early intervention for victims of domestic abuse up to 3 years before a potential police PPN submission. The findings demonstrate several risk factors that reflect vulnerability; when the perpetrator exhibited a pre-existing pre-disposition for violence, either in hurting others or with a history of violence, the domestic abuse victims were found to have an increased vulnerability for future outcomes. Furthermore, pregnant victims showed heightened vulnerability and had poorer outcomes; research has indicated that domestic abuse has been independently associated with the birth of a low-birthweight baby.[19]

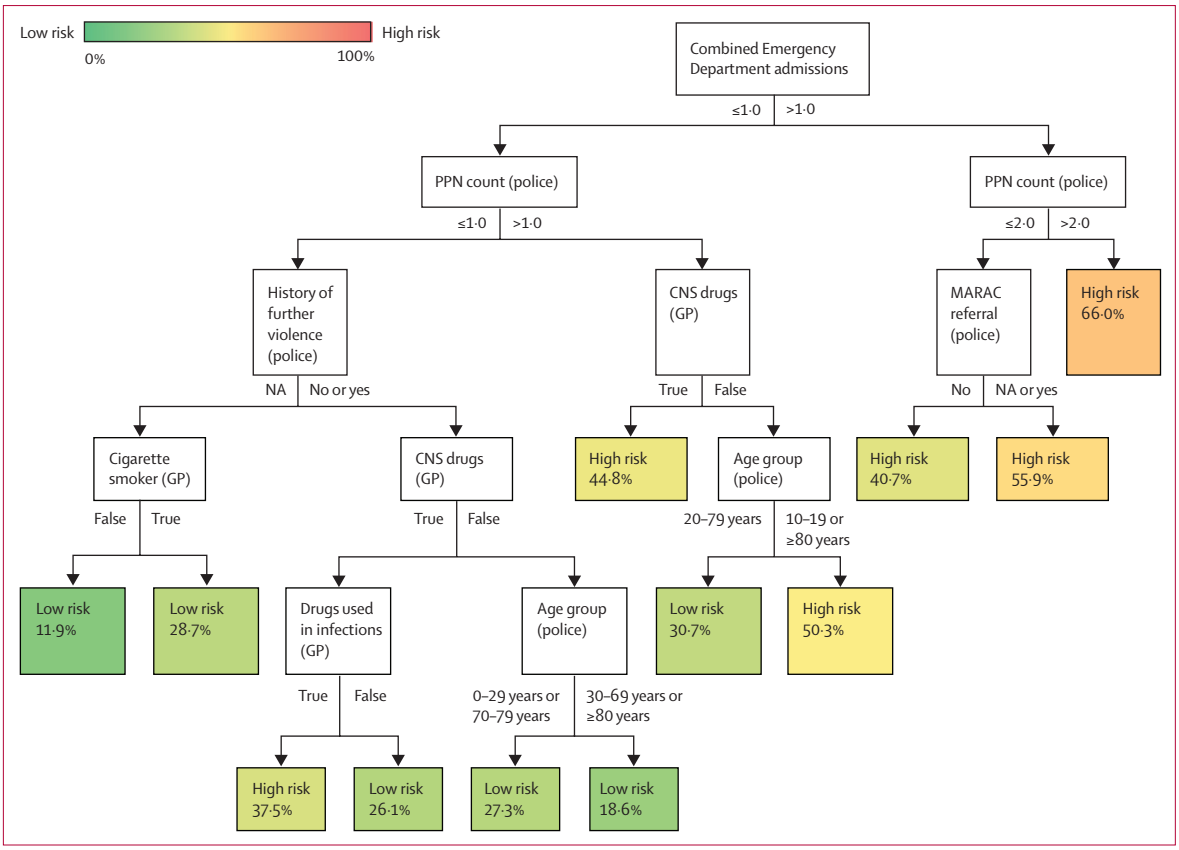


Figure 2: **Decision tree to examine clustering of those who experience an outcome using data from other sources up to 1 year before the index PPN submission date**

GP=general practitioner. MARAC=multi-agency risk assessment conference. NA=not applicable. PPN=Public Protection Notification.

Additionally, victims were shown to be known to Emergency Departments before the police are involved. Highly vulnerable individuals might interact frequently with Emergency Department health-care services but remain unknown to the police until a domestic abuse incident that requires an intervention;[4,11–13] the most vulnerable individuals have been identified as those who had more than seven Emergency Department admissions up to 3 years before the index PPN submission.

The decision tree analysis highlighted several risk profiles with differing clusters of risk factors that might reflect different pathways through police contact, health care, and individual contexts. Our findings corroborated previous observations that the highest risk group of people experiencing domestic abuse are those well known to Emergency Department health-care services.[11] In our decision tree analysis, we found that an additional indicator for high risk was having had interactions with their GP, either for smoking cessation advice or by being identified as a cigarette smoker. This observation is in line with research that indicated that domestic abuse is associated with adverse health behaviours, such as substance abuse and addiction.[4,7–9,11,13] Smoking could be indicative of an addiction behaviour that might serve as a coping mechanism.[20] Furthermore, smoking behaviour has been shown to be more prevalent in individuals from an economically deprived background,[21–23] which has been found to be associated with domestic abuse within the household.[6]

A further set of factors suggested a subset of victims that are at high risk for a major health outcome. These individuals are repeatedly in contact with police after the index PPN submission and before the subsequent outcome event. Both decision trees indicated that those at the highest risk are split by the presence of more than two police visits. Individuals with less than two police visits but with a MARAC referral were also considered to be at high risk. This finding might indicate a progressively worsening home situation. These findings are similar to what has been found in victims of intimate partner violence; a study in the USA observed that Emergency Department use was associated with an increased number of police calls,[13] whereas another study in Canada found it was associated with more violent abuse than for those who do not use the Emergency Department.[4] These findings indicate that the manifestation of domestic abuse, specifically intimate partner violence in these studies, has shifted from a situation of coercive control to physical violence.

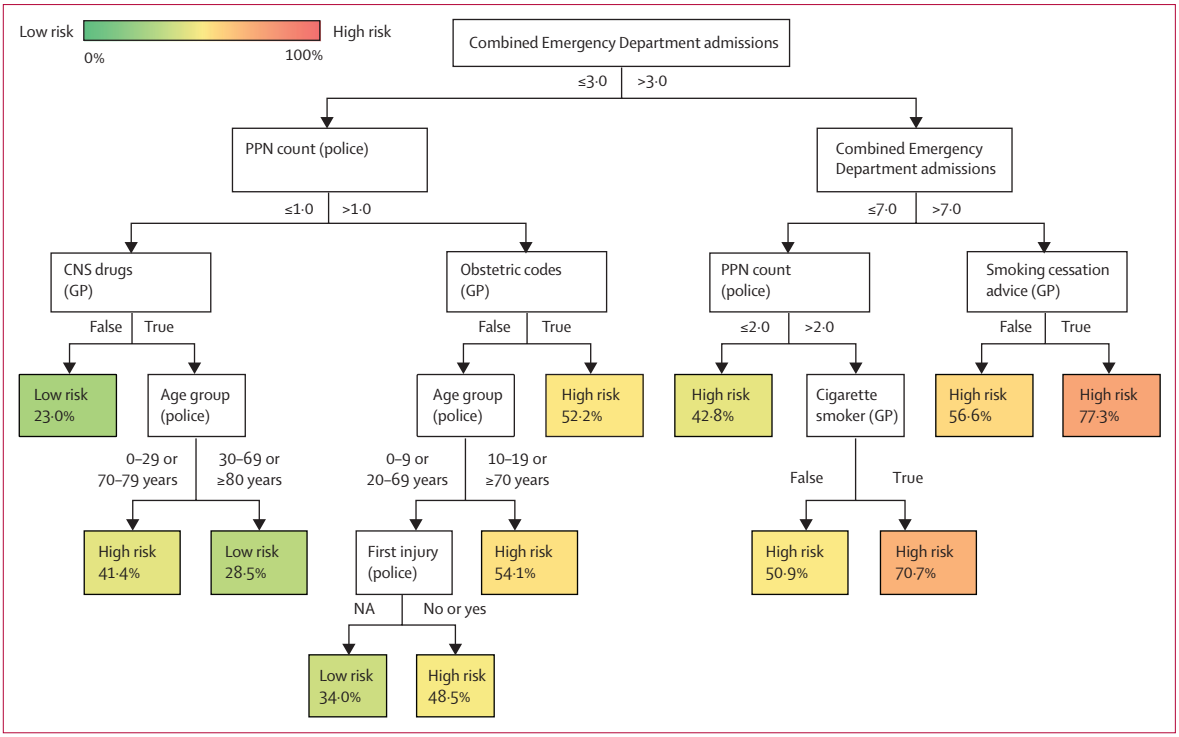


Figure 3: **Decision tree to examine clustering of those who experience an outcome using data from other sources up to 3 years before the index PPN submission date**

GP=general practitioner. NA=not applicable. PPN=Public Protection Notification.

In contrast another group comprised individuals who were less often in contact with Emergency Department health-care services. This group consists of people who have had more than one police visit after the domestic abuse incident, with less than three Emergency Department admissions up to 3 years before, or less than one Emergency Department admission up to 1 year before the domestic abuse incident. Obstetric codes recorded by the victim's GP up to 3 years before were indicative of those at highest risk within this group: mothers with young children. Several studies have observed that victims of intimate partner violence, which is included in the definition of domestic abuse by the DASH questionnaire, are more likely to be in their childbearing years;[11,24–26] those who suffer abuse during pregnancy reported that it worsened throughout the duration, whereas for some who previously suffered abuse, they indicated that it was a protective period.[25] Research has indicated that having a child can increase economic pressures;[27] economic hardship increases stress on relationships, which has been shown to escalate the incidence of abuse.[27] The survival analysis also highlighted that victims who are pregnant at the time of PPN submission were more likely to have an outcome in the following year. Conversely, cases characterised by conflict over child contact had a decreased risk of experiencing a future outcome, which might be interpreted that a co-parent and potential perpetrator no longer cohabits within the same household, which could decrease exposure to abusive behaviour.

Finally, we identified a group that was relatively unknown to both Emergency Department health-care services and police, but who nonetheless could be considered vulnerable individuals from the presence of GP codes. Those who had less than one police visit after the PPN submission were more likely to experience an outcome if they have been prescribed CNS drugs and were in the extreme age groups of 0–29 years or 70–79 years, compared with those who have not been prescribed these drugs or who are in different age groups. Furthermore, victims who were questioned by the police about the perpetrator's violent tendencies at the incident and were prescribed both CNS drugs and drugs used in infections are more likely to experience the outcome compared to those who are not prescribed such drugs. These codes could suggest a stressful home environment, which is associated with adverse mental health symptoms. The presence of infection codes might indicate that these people have injuries or poor living conditions.

Our findings suggest that people who are attended to by the police for cases of abuse in the household have previously been in contact with various service providers, including GPs and Emergency Department health care. The linking of data from different organisations might enhance the efficiency and effectiveness of these

organisations' response systems. By using anonymised data sharing and linkage across multiple agencies we could identify warning signs, such as frequent visits to an Emergency Department. It could also help refine the DASH questionnaire adding additional risk factors that have been identified from linked data research, such as questions regarding previous Emergency Department admissions and the prescription of antidepressants or antibiotics from their GP. The use of an anonymised linkage system with a trusted third party enables research and system learning while preserving confidentiality and anonymity of the individuals. However, the implementation of cross-organisational data linkage at the national level for long-term purposes beyond research to identify individuals requires consultation with the public and consensus on the appropriate data types, linkage objectives, and purpose for linkage. Consensus is necessary to ensure that data sharing between organisations does not prevent individuals seeking help from Emergency Departments or similar services due to fear of being identified by the police, as such awareness has potential to deter help-seeking behaviours. The results of this study underscore the importance of providing training for Emergency Department personnel to recognise and address potential cases of abuse.

There are several limitations to this study. The PPN dataset comprises records spanning 2015–20 as PPN records were not available before 2015. Consequently, left-censoring arose in this study. In addition, the inclusion criteria required that the victim first has a PPN submission before experiencing a domestic abuse-related outcome. As such, those individuals who are experiencing domestic abuse without engaging with the police are excluded from this analysis. Furthermore, the index PPN submission within this study might not be the actual first PPN submission; thus, the actual time to event might be different to the calculated time to event between the index PPN and the outcome event. Moreover, this study is predominantly descriptive; therefore factors identified cannot be used to establish cause and effect and residual confounding might operate. Finally, restricting the analysis to non-obstetric Emergency Department admissions for women might have resulted in an underestimation of the influence of household abuse in pregnancy. Notably, some events, such as pre-term delivery, which could be triggered by abuse, would have been excluded.

Highly vulnerable individuals frequently interact with health-care services but remain unknown to the police before an incident that requires an intervention. This finding underscores the potential value of linking data across different agencies to facilitate targeted prevention measures instead of reactive ones. Moreover, identifying individuals at high risk following a police interaction could enable the establishment of protective measures. The data generated by this study have the potential to identify risk without relying on disclosure, forming a foundation for enhanced integration. Further research is needed to validate these findings and examine whether early identification of these groups by the police or in the health-care setting could lead to improved outcomes, in addition to examining the outcomes stratified by demographic profiles.

Contributors

MAB, AJ, KHa, SM, and SB conceptualised the study. AB and SM designed the study. SB secured funding, contributed to the methodology, and undertook the role of project principal investigator. NK, TLW, and CM collected the data, with NK and TLW directly accessing and verifying the underlying data. NK, TLW, AB, and JK undertook the data analysis. NK, JK, and SB interpreted the data. NK and JK generated the figures. NK drafted the report. BR, CM, JE, KHu, MAB, AJ, SM, and SB all undertook project supervisory and advisory roles for their specialties. KHa acted as a patient and public involvement advisor. All authors contributed to revisions, approved the final version, and were responsible for the decision to submit the manuscript for publication.

Declaration of interests

We declare no competing interests.

Data sharing

The data that support the findings of this study are available from the Secure Anonymised Information Linkage (SAIL) databank,[15] but restrictions apply to the availability of these data, which were used under licence for the current study and so are not publicly available. Interested individuals can apply to the SAIL databank for access, and once approved, can apply to the corresponding author.

Acknowledgments

This research was funded by the National Institute for Health Research, Public Health Research Board (reference number NIHR133680: Unlocking Data to Inform Public Health Policy and Practice). The study was also supported by Health Care Research Wales through the National Centre for Population Health and Wellbeing Research, supported by ESRC through Administrative Data Research Wales, and received infrastructure support through Health Data Research UK. This study makes use of anonymised data held in the SAIL databank.[15] We would like to acknowledge all the data providers who make anonymised data available for research.

References

1 His Majesty's Inspectorate of Constabulary and Fire & Rescue Services. Police effectiveness 2015 (vulnerability). Dec 15, 2015. https://www.justiceinspectorates.gov.uk/hmicfrs/publications/police-effectiveness-vulnerability-2015-south-wales/ (accessed April 17, 2023).

2 Richards L. Domestic Abuse, Stalking and Harassment and Honour Based Violence (DASH, 2009) risk identification and assessment and management. 2009. https://proceduresonline.com/trixcms/media/6627/dash-risk-assessment.pdf (accessed April 17, 2023).

3 UK Parliament. Domestic Abuse Act. 2021. https://www.legislation.gov.uk/ukpga/2021/17/part/1/enacted (accessed April 17, 2023).

4 Nesca M, Au W, Turnbull L, Brownell M, Brownridge DA, Urquia ML. Intentional injury and violent death after intimate partner violence. A retrospective matched-cohort study. *Prev Med* 2021; **149:** 106616.

5 Sardinha L, Maheu-Giroux M, Stöckl H, Meyer SR, García-Moreno C. Global, regional, and national prevalence estimates of physical or sexual, or both, intimate partner violence against women in 2018. *Lancet* 2022; **399:** 803–13.

6 Nesset MB, Gudde CB, Mentzoni GE, Palmstierna T. Intimate partner violence during COVID-19 lockdown in Norway: the increase of police reports. *BMC Public Health* 2021; **21:** 2292.

7 Bacchus LJ, Ranganathan M, Watts C, Devries K. Recent intimate partner violence against women and health: a systematic review and meta-analysis of cohort studies. *BMJ Open* 2018; **8:** e019995.

8 Chandan JS, Thomas T, Bradbury-Jones C, et al. Female survivors of intimate partner violence and risk of depression, anxiety and serious mental illness. *Br J Psychiatry* 2020; **217:** 562–67.

9 Hegadoren KM, Lasiuk GC, Coupland NJ. Posttraumatic stress disorder part III: health effects of interpersonal violence among women. *Perspect Psychiatr Care* 2006; **42:** 163–73.

10 Alejo K. Long-term physical and mental health effects of domestic violence. May 1, 2014. https://scholarworks.sjsu.edu/themis/vol2/iss1/5 (accessed Feb 28, 2022).

11 Kothari CL, Rhodes KV. Missed opportunities: emergency department visits by police-identified victims of intimate partner violence. *Ann Emerg Med* 2006; **47:** 190–99.

12 Kothari CL, Rhodes KV, Wiley JA, et al. Protection orders protect against assault and injury: a longitudinal study of police-involved women victims of intimate partner violence. *J Interpers Violence* 2012; **27:** 2845–68.

13 Rhodes KV, Kothari CL, Dichter M, Cerulli C, Wiley J, Marcus S. Intimate partner violence identification and response: time for a change in strategy. *J Gen Intern Med* 2011; **26:** 894–99.

14 Coker AL, Smith PH, Bethea L, King MR, McKeown RE. Physical health consequences of physical and psychological intimate partner violence. *Arch Fam Med* 2000; **9:** 451–57.

15 Campbell JC. Health consequences of intimate partner violence. *Lancet* 2002; **359:** 1331–36.

16 Lyons RA, Ford DV, Moore L, Rodgers SE. Use of data linkage to measure the population health effect of non-health-care interventions. *Lancet* 2014; **383:** 1517–19.

17 Jones KH, Ford DV, Thompson S, Lyons RA. A profile of the SAIL databank on the UK Secure Research Platform. *Int J Popul Data Sci* 2019; **4:** 1134.

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

19 Bandyopadhyay A, Jones H, Parker M, et al. Weighting of risk factors for low birth weight: a linked routine data cohort study in Wales, UK. *BMJ Open* 2023; **13:** e063836.

20 Sullivan TP, Flanagan JC, Dudley DN, Holt LJ, Mazure CM, McKee SA. Correlates of smoking status among women experiencing intimate partner violence: substance use, posttraumatic stress, and coping. *Am J Addict* 2015; **24:** 546–53.

21 Martire KA, Clare P, Courtney RJ, et al. Smoking and finances: baseline characteristics of low income daily smokers in the FISCALS cohort. *Int J Equity Health* 2017; **16:** 157.

22 Guillaumier A, Twyman L, Paul C, Siahpush M, Palazzi K, Bonevski B. Financial stress and smoking within a large sample of socially disadvantaged Australians. *Int J Environ Res Public Health* 2017; **14:** 231.

23 Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol* 2017; **186:** 1026–34.

24 Ribeiro MRC, Batista RFL, Schraiber LB, et al. Recurrent violence, violence with complications, and intimate partner violence against pregnant women and breastfeeding duration. *J Womens Health (Larchmt)* 2021; **30:** 979–89.

25 Alhusen JL, Ray E, Sharps P, Bullock L. Intimate partner violence during pregnancy: maternal and neonatal outcomes. *J Womens Health (Larchmt)* 2015; **24:** 100–06.

26 MacMillan HL, Wathen CN. Children's exposure to intimate partner violence. *Child Adolesc Psychiatr Clin N Am* 2014; **23:** 295–308, viii–ix.

27 Friedline T, Chen Z, Morrow S. Families' financial stress & well-being: the importance of the economy and economic environments. *J Fam Econ Issues* 2021; **42** (suppl 1): 34–51.

28 Spooner M. Does eligibility for protection orders prevent repeat abuse of domestic abuse victims in Caribbean states? *J Fam Violence* 2009; **24:** 377–87.

My input

I contributed as a co-project investigator for a successful National Institute for Health Research (NIHR133680) grant, which provided the necessary funding and resources to conduct this important research. This grant enabled the integration of police data into the WECC Phase 4 framework, significantly enhancing our ability to analyse and understand the dynamics of DA.

I developed the study design and contributed to the methodology used to explore the research objectives. I also provided supervisory guidance on data linkage and preparation, ensuring the accuracy and integrity of the data used in this study. Additionally, I worked on the validation of the scripts developed for the modelling of the data and contributed to the development of the manuscript.

Impact

- This paper was published in *The Lancet Public Health* journal in 2023, one of the most prestigious journals in public health research.
- This work has been cited by nine other published studies, highlighting the paper's importance and relevance in the research field.
- This work has laid the foundation for future research grants aimed at investigating linked police and administrative data.

Conclusion

The findings from this study reinforce the critical need for an integrated approach to addressing DA and its consequences on children, as they are at higher risk of experiencing adverse health outcomes due to DA. DA substantially impacts younger individuals (increasing the risk of adverse physical and mental health outcomes) and pregnant women's (increasing the risk of preterm birth, LBW and postnatal mental and physical health issues), further heightening the vulnerability of these groups to severe outcomes. This linked police and healthcare data study provides a clearer picture of risk factors and outcomes, emphasising the importance of early intervention and coordinated support services. While the focus here is on family-level vulnerabilities, such as exposure to DA, the following chapter will explore the association between family environment and the health conditions of children who are at a higher risk of early alcohol use, one of the most severe risk-taking behaviours that can develop during adolescence.

Chapter 6: Health and household environment factors linked with early alcohol use in adolescence: a record-linked, data-driven, longitudinal cohort study

Critical summary

Background

Early onset of alcohol use is a predictor of adverse outcomes, increasing a child's vulnerability later in life, including low academic achievement, poor physical and mental health outcomes and a higher risk of substance abuse (81–83). Existing research on the risk factors associated with early alcohol use has primarily focused on family characteristics and individual socio-economic, neurocognitive, behavioural, or emotional factors, either individually or in combination (84–86). However, the relationship between a child's health status and subsequent alcohol use remains underexplored, with limited data-driven exploration of these risk factors in the literature. This study integrates hypothesis-based knowledge with data-driven insights to investigate factors associated with early alcohol use. Employing a two-stage data-driven approach, the research assesses the interplay between childhood health factors, household environments and alcohol-related outcomes in adolescence. The study aims to provide a new methodological framework for investigating the risk factors associated with early alcohol use, thereby enhancing current understanding of this critical public health issue.

Utilisation of administrative and other data

One of the novelties of the current study is hybridisation, which involves combining survey with routine data. This hybridisation is a powerful approach in data science, bringing together the strengths of both survey and routine data sources that complement each other. The MCS is a survey data that provides detailed child and household-level information at various stages of the child's life, while the longitudinal nature of the routine data adds significant richness to the study. To address early-life vulnerabilities effectively, having a holistic picture of the household, along with the health of the family and child, is crucial which was offered by MCS survey data. Hybridisation enables the establishment of a comprehensive framework for investigating early alcohol use

This study employs a two-stage data-driven approach. In stage one, longitudinal survey data from the MCS (87) were utilised, then linked to participants' primary care electronic health records (EHRs) from WLGP and secondary care EHRs from PEDW. The objective of linking the longitudinal survey cohort with administrative EHR data includes a) leveraging 11 years of data from the MCS, encompassing self-reported alcohol use, parent-reported family-level data and child-specific information; and b) integrating this

with 10 years of longitudinally followed EHR data (one year prior to the collection of alcohol data) in the form of ICD-10 and READ codes Version 2.

Stage one provided a list of significant risk factors for early alcohol use obtained from linked survey and routine data collection. These findings informed the development of analogous risk factors in stage two using solely linked routine data. While stage one developed a risk profile based on survey participants, stage two focused on constructing a risk profile for the entire population. The extensive EHRs, encompassing ICD-10 and READ codes, provided a rich resource for a comprehensive analysis of health patterns and other risk factors associated with early alcohol use. This novel two-stage approach facilitated the creation of a comprehensive dataset that combined the strengths of both survey and routine data, helping to reduce the limitations of the routine data.

Application of data science methods

In stage one, a chi-square (χ^2) feature selection method was employed to identify the most significant health codes from 10 years of EHR data. Following this, a stepwise LR model was developed to incorporate the interdependence between explanatory variables derived from both survey and routine data. This process identified the most statistically significant variables associated with alcohol use outcomes, helping to reduce the variable space and optimise the time required to recreate analogous variables for stage two.

In stage two, analogous variables were constructed from the routine data, as detailed in the paper. This approach optimised the utilisation of routine data in the analysis. The hybridisation of diverse data types, as implemented in this study, represents a novel approach that merges the complementary strengths of EHRs with the personal insights derived from questionnaire-based cohort data. This integration provides a robust foundation for findings generalisable to the broader population.

This methodological approach strengthens the analysis by systematically identifying the most relevant factors contributing to the risk of early alcohol use. The data-driven approach eliminates the risk of overlooking any critical variables that influence alcohol-related outcomes.

Early-life vulnerability profiling

The findings of the study reveal the critical role of both individual health status and family dynamics in predicting early alcohol use. The longitudinal nature of the study allowed for an exploration of changes in exposure during the first 11 years of life, providing a clearer understanding of how early experiences shape later alcohol use. The use of survey data adds value by capturing in-depth family-level circumstances that contribute to the development of risk profiles in early life helping to predict adverse outcomes. This combination improves our understanding of health issues and strengthens the overall analysis.

The record-linked, data-driven methodology demonstrates the power and effectiveness of integrating survey data with EHRs to build a comprehensive risk profile, applicable to other areas of vulnerability profiling. The study's findings (from both stages) also highlight that children who receive support for their health needs, such as healthcare records for vaccinations, attendance at routine health examinations with their GP and contact with health services recorded in primary and secondary care are at a lower risk of early alcohol use. This suggests that a lack of regular healthcare contact may serve as a proxy for increased risk. By enabling precise vulnerability profiling, this rigorous methodological framework contributes to the development of more targeted public health interventions and strategies to reduce early alcohol use among at-risk populations.

Published journal paper

International Journal of Population Data Science

Journal Website: www.ijpds.org

Health and household environment factors linked with early alcohol use in adolescence: a record-linked, data-driven, longitudinal cohort study

Amrita Bandyopadhyay[1,2], Sinead Brophy[1,2,3], Ashley Akbari[1,3], Joanne Demmler[3], Jonathan Kennedy[2], Shantini Paranjothy[4,5], Ronan A. Lyons[1,2,3], and Simon Moore[4,6,*]

Submission History

Submitted:	16/11/2021
Accepted:	06/05/2022
Published:	07/07/2022

[1]Administrative Data Research Wales, Swansea University Medical School, Wales SA2 8PP, UK
[2]National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Wales, SA2 8PP, UK
[3]Health Data Research UK, Swansea University Medical School, Wales, SA2 8PP, UK
[4]School of Dentistry, Cardiff University, Cardiff, Wales, CF14 4XY, UK
[5]University of Aberdeen, Aberdeen Health Data Science Centre, Institute of Applied Health Sciences, Polwarth Building, Foresterhill, Aberdeen, AB25 2ZD
[6]Security, Crime and Intelligence University Innovation Institute, Social Science Research Park, Cardiff University, Maindy Road, Cardiff, CF24 4HQ, UK

Abstract

Introduction
Early alcohol use has significant association with poor health outcomes. Individual risk factors around early alcohol use have been identified, but a holistic, data-driven investigation into health and household environmental factors on early alcohol use is yet to be undertaken.

Objectives
This study aims to investigate the relationship between preceding health events, household exposures and early alcohol use during adolescence using a two-stage data-driven approach.

Methods
In stage one, a study population (N = 1,072) were derived from the Millennium Cohort Study (MCS) Wales (born between 2000–2002). MCS data were first linked with electronic-health records. Factors associated with early (<= eleven years old) alcohol use were identified using feature selection and stepwise logistic regression. In stage two, analogous risk factors from MCS were recreated for whole population (N = 59,231) of children (born between 1998-2002 in the Welsh Demographic Service Dataset) using routine data to predict the alcohol-related health events in hospital or GP records.

Results
Significant risk factors from stage two included poor maternal mental (adjusted odds ratio [aOR] = 1.31) and physical health (aOR = 1.25), living with someone with alcohol-related problem (aOR = 2.16), single-adult household (aOR = 1.45), ever in deprivation (aOR = 1.66), child's high hyperactivity (aOR = 3.57), and conduct disorder (aOR = 3.26). Children with health events, whose health needs are supported (e.g., are taken to the doctor), are at lower risk of early alcohol use.

Conclusion
Health events of the family members and the child can act as modifiable exposures and may therefore inform the development of prevention initiatives. Families with known alcohol problems, living in deprivation, experiencing child behavioural problems and those who are not taken to the doctor are at higher risk of early drinking behaviour and should be prioritised for early years support and interventions to target problem drinking in young people.

Keywords
alcohol; adolescent; data linkage; electronic health records (EHRS); cohort study

*Corresponding Author:
Email Address: [redacted] (Simon Moore)

https://doi.org/10.23889/ijpds.v7i1.1717

Introduction

Alcohol use in childhood is associated with the risk of later alcohol abuse, alcohol dependence [1] and several negative outcomes including poor educational achievement, death and disability [2–5]. Known factors that predict early alcohol use include a child's hyperactivity and conduct disorder [6, 7], lack of family support, household dysfunction, parental alcohol drinking pattern, parental indifference towards young persons' alcohol use [8–11] and adverse childhood experiences (ACEs) (e.g., child abuse and parental discord) [12]. Current research has largely focused on the family environment, individual level socio-demographic, neurocognitive, behavioural or emotional features, individually or in combination [13–15]. Although it is known that ACEs have a detrimental impact on a child's health in early life [16, 17], it is not known whether a child's own health status is associated with subsequent alcohol use and alcohol-related health outcomes.

Child health is a broad term that includes maintaining and protecting physical, mental and social health [18]. Broadly, there are two dominant methodological approaches in the investigation of child alcohol use that are increasingly regarded as complementary [19]. First, survey methodology allows researchers to focus on specific exposures and outcomes, such as volume of alcohol consumed, and to tailor validated [20] instruments to address preconceived study hypothesis [2]. Limitations include relatively small sample size, non-response, selection and volunteer bias [21]. Second, the analysis of routinely collected electronic health records (EHRs) facilitates the inclusion of a greater number of individuals, even entire populations, than is feasible using surveys. The analysis of whole population EHRs, however, imposes challenges relating to the processing and management of data, including addressing missing data on informative variables [22]. For example, EHRs are unlikely to capture occasional alcohol consumption but would be expected to capture health outcomes relating to hazardous alcohol use.

Existing literature on this topic has predominantly focused on preconceived study hypothesis [2], however this increases the chance of missing risk factors which have not already been identified. In contrast to this, a data-driven framework would avoid the limits of a pre-defined and hypothesis-bound investigation and significantly open up the exploration of the variable space. We anticipate that this will provide new insights and will ultimately help to develop a better understanding of the research problem under investigation. Hence, the current study does not focus on an explicit causal analysis, rather we aim to merge hypothesis-based knowledge with data-driven insights to investigate the risk factors associated with early alcohol use.

In this study we assess the relationship between childhood health factors, household environment and alcohol-related outcomes during adolescence using a two-stage data-driven approach. These broad categories of risk factors were based on hypothesis-based knowledge as discussed above. This method brings together a hypothesis-based study design followed by a data-driven approach which complements and minimises the limitation of both study designs.

Methods

A two-stage data-driven approach has been undertaken to investigate the association between the specific risk factors and the outcome in this study. In stage one, a machine learning feature selection algorithm and a classifier were used to identify the health conditions and socio-demographic factors associated with early alcohol use from linked EHRs and Millennium Cohort Study (MCS) survey data. In stage two, analogous risk factors identified from stage one were then sought in routine data and an analytic approach was used to determine the prediction model. The linked routinely collected EHRs and vast volume of administrative data from the whole population of Wales was analysed to determine the effect of the risk factors identified in the MCS data analysis as predictors to target alcohol-related health outcomes in the general adolescent population.

Stage one – Millennium Cohort Study (MCS)

Participants

The MCS is a longitudinal birth cohort of children born in the UK between the years 2000 and 2002 [23]. Parents of the original 18,819 singleton children were interviewed from all parts of UK when their child was nine months old, of those 1,951 were interviewed in Wales. Subsequent interviews took place at ages three, five, seven and eleven years of age. Written consent to link MCS children with their routine EHRs up to age fourteen years was obtained from their parents at the interview undertaken when children were seven years of age. Data of the 1,838 consented singleton children resident in Wales was subsequently linked with their EHRs. The study population included children who also participated in the interview at age eleven years, as the primary outcome data were collected at that point. The current study excluded participants who did not have a general practitioner (GP) record in the Welsh Longitudinal General Practice (WLGP) dataset before they were eleven years of age (Supplementary Figure 1).

Exposure

The study included parent reported socio-demographic and family-related variables for children from MCS interviews which took place between the age of nine months and seven years of the children. These include child's sex, mother's socio-economic classification (SEC), household poverty level (whether the household income was above/below 60% of national median using a modified Organisation for Economic Co-operation and Development scale), living area (based on 2005 Rural/Urban Area Classification), mother's alcohol use during and post pregnancy, lone parent carer, and number of children. Based on lone parent status, the total number of siblings at household and total number of household members, the study derived a binary variable to identify whether the child was residing with any other additional household members. Using both parents' responses on alcohol consumption, guardian alcohol use variables were derived. Children's emotional and behavioural difficulties were measured using the parent completed Strength and Difficulty Questionnaire (SDQ) [24]. Since most of these variables are time varying (and collected from MCS at ages nine months

until age eleven years) aggregated summary variables were derived based on average values. These variables include SDQ, mother's SEC, lone parent status, guardians' alcohol use, living area, poverty indicator, additional household member and mother's alcohol use after their child was born. The exposure variables from MCS have been described in Table 1.

The health records of the children were also considered as the exposures for risk of early alcohol use. EHRs of the MCS children obtained from hospital admission record and primary care events within the Patient Episode Database for Wales (PEDW) and the WLGP dataset. A broad list of explanatory health codes was constructed using the three-digit ICD-10 codes and Read Code Version 2 recorded in PEDW and WLGP from birth until age ten (one year before the alcohol data were collected). Wales Electronic Cohort for Children (WECC) [25] containing further details on child health in Wales, were used to obtain age and maternal age at birth.

Outcome

Alcohol data for MCS children were obtained from a self-report questionnaire at age eleven (Supplementary Table 1). Based on the responses to the questionnaire the children were classified into two groups: those who had consumed alcohol (case) and those who had not (non-case). Those who did not answer or provided contradictory responses were removed from analyses (Supplementary Figure 1).

Statistical analysis

In the cohort exposure dataset, the participants with more than 10 missing variables (out of 13) were removed from analyses to ensure the accuracy of the data. An explanatory variable with less than 10% missing data had been imputed using a predictive mean matching (PMM) imputation method [26, 27].

To identify the health codes that were associated with early alcohol use from the large volume of linked EHRs spanning 10 years, a chi-square (χ^2) feature selection method was applied [28]. A critical threshold value $\chi^2 \geq 2.706$ (one degree of freedom, $p \leq 0.1$) was applied and health codes with a χ^2 above this threshold were retained in subsequent analyses. A multivariate stepwise logistic regression with bidirectional (forward and backward) search was then performed for the exposure variables to obtain the best-fit model [29]. In stepwise model the variables with least significance were removed at each iteration step and the final model was selected based on the minimum Akaike Information Criterion (AIC) value. From the final model, only the statistically significant ($p \leq 0.05$) variables were selected as significant predictors associated with the risk of early alcohol use leading to a further reduction in variable space. This is justified due to the following reasons.

- The variable selection process facilitates the choice of best model by incorporating the interdependence between the explanatory variables.

- The approach only considers the statistically significant variables for the stage two analysis which reduces the variable space and optimises the time to recreate analogous variables.

Stage two – whole population

Participants

All children born between 1st January 1998 and 31st December 2002 and were resident in Wales during the first fourteen years of their life were included in the whole population dataset. The study population was selected from the Welsh Demographic Service Dataset (WDSD), which is an administrative dataset of individuals living in Wales registered with a GP. The participants without continuous record in the WLGP from age six months to fourteen years were excluded to ensure a complete follow-up period.

Exposure

Analogous risk factors to those identified in the MCS analysis were created using the WDSD, WLGP and PEDW data. The study used an encrypted household identifier known as residential anonymised linking field (RALF) which enabled the participants to be linked with other household members and related records [30]. Each RALF is associated with the smallest geographical representation known as lower super output area (LSOA) which again is associated with a Welsh Index of Multiple Deprivation (WIMD) rank aggregated into a quintile or decile scale. Overall and employment WIMD scores were used as the measure of deprivation from routine data in the study. The main explanatory variables derived from routine data for the whole population analysis include child's sex, employment deprivation and overall deprivation, living with single adult, mother's alcohol-related condition during pregnancy, living with household member with alcohol-related condition, living area, maternal age, gestational age, and child mental and physical health. To be consistent with the MCS data, primary exposure data were collected for children up to age seven years. For time varying variables, the study used the same time points as MCS (birth to nine months, nine months to three years, three to five years, and five to seven years) and derived aggregated summary variables for the risk factors. Detailed descriptions of the variables are available in Supplementary Table 2.

Outcome

Alcohol-related health events across the whole population cohort were obtained from ICD-10 codes in PEDW (Supplementary Table 3) and Read codes in WLGP (Supplementary Table 4) between the age seven and fourteen years [31].

Statistical analysis

As the case (alcohol-related EHRs) to non-case (no alcohol-related EHRs) ratio was 1:99 in the whole population cohort and unbalanced, to improve the efficiency and the sensitivity of model performance case-control selection was undertaken by randomly selecting 20 non-cases for each sex matched case [32]. The dataset was randomly split into a training (70%) and test set (30%). Logistic regression was used to obtain the best-fit model on the training data. Model prediction on the test data provided a predictive probability of the expected outcome associated with each individual. Model prediction

Table 1: Socio-demographic characteristics of the MCS population (following imputation) and whole population sample with descriptive statistics

MCS			**Whole Population**		
Child Sex	n	%		n	%
Female	521	48.60	Female	28,770	48.57
Male	551	51.40	Male	30,461	51.43
Deprivation					
Mother Socio economic classification (SEC)			Overall deprivation		
Always managerial or intermediate	377	35.17	Low (WIMD quintile >=3)	29,102	49.13
Always semi-employed, self-employed, semi-routine or routine	280	26.12	High (WIMD quintile <3)	24,701	41.70
Unknown	415	38.71	Borderline (ever belong to high group but not always)	5,428	9.16
Poverty indicator			Employment deprivation		
Above poverty level	539	50.28	Low (WIMD quintile >= 3)	29,394	49.63
Below poverty level	270	25.19	High (WIMD quintile < 3)	24,774	41.83
Ever been below poverty level	263	24.53	Borderline (ever belong to high group but not always)	5,063	8.55
Household alcohol use					
Mother's alcohol use during pregnancy			Mother's alcohol-related health condition during pregnancy		
Never	752	70.15	No	55,251	93.28
Low (less than once a month or 1–2 times a month)	218	20.34	Yes	3,980	6.72
High (more than 1–2 times a month)	102	9.51			
Mother's alcohol use after child was born					
Never	82	7.65			
Low	500	46.64			
High	490	45.71			
Guardian alcohol use			Household member identified with alcohol-related hospital admission		
Low	247	23.04	No	57,799	97.58
Moderate	524	48.88	Yes	1,432	2.42
High	233	21.74			
Variable	68	6.34			
Living area					
Rural	238	22.20		14,760	24.92
Urban	779	72.67		41,907	70.75
Ever been urban	55	5.13		2,564	4.33
Maternal age at child's birth					
Less than 20 years	102	9.51		7,111	12.01
20 to 24 years	202	18.84		9,266	15.64
25 to 29 years	305	28.45		17,389	29.36
30 to 34 years	324	30.22		17,005	28.71
35 years and over	139	12.97		8,460	14.28
Gestational age					
Not term	52	4.85		1,317	2.22
Term	1,020	95.15		57,914	97.78
Household composition					
Siblings at home			Living with single adult		
No sibling	129	12.03	No	33,662	56.83
One sibling always or at some point	493	45.99	Yes	8,425	14.22
More than one sibling ever	450	41.98	Ever been	17,144	28.94

(Continued).

Table 1: Continued

MCS			Whole Population		
Lone parent					
No	754	70.34			
Yes	130	12.13			
Ever been	188	17.54			
Additional household member					
No	792	73.88			
Yes	118	11.01			
Ever had	162	15.11			
Mother's health					
Longstanding illness			Mother's any comorbidity		
No	589	54.94	No	46,170	77.95
Yes	170	15.86	Yes	13,061	22.05
Varies	313	29.20	Mother's psychosis disorder		
			No	58,924	99.48
			Yes	307	0.52
			Mother's common mental health condition		
			No	28,603	48.29
			Yes	30,628	51.71

Table 2: Health codes identified as risk factors for early alcohol use by chi-square feature selection method in the MCS cohort and the percent of sample with these codes present in whole population (WP) following selection

Health code	Description of the code	Type of code	chi-square	MCS (%)	WP (%)
Read code H05%	Upper respiratory infections	Diagnosis	.60	62.50	59.95
Read code K2%	Male genital organ diseases	Diagnosis	7.77	12.41	8.46
Read code 919%	Child health surveillance related administrative code	Administrative	6.07	25.56	30.70
Read code 64N%	Child physical health examination	Administrative	4.63	17.35	15.56
Read code 656%	Tetanus vaccination	Administrative	4.11	28.26	34.21
ICD-10 code Z%	Factors influencing health status and contact with health services	Diagnosis	3.90	27.99	20.68
Read code 654%	Diphtheria vaccination	Administrative	3.69	27.71	-
Read code 655%	Pertussis vaccination	Administrative	3.35	29.94	-
Read code F%	Nervous system and/or sense organ diseases	Diagnosis	3.04	70.24	-
Read code F4%	Disorders of eye and adnexa	Diagnosis	3.00	46.27	-
Read code K27%	Disorders of penis	Diagnosis	2.99	9.42	–
Read code etc.%	Trimethoprim, an antibiotic used mainly in the treatment of bladder infections	Medication	2.93	16.70	–
Read code 4%	Laboratory test and procedures (e.g. urine culture, blood test)	Administrative	2.89	60.73	–

codes were not selected by the logistic regression models, hence were not selected for WP analysis

was quantified by performance accuracy, sensitivity, specificity, positive predictive value, and negative predictive value.

MCS and routine EHRs were anonymously linked and accessed within the Secure Anonymised Information Linkage (SAIL) Databank. Linkage was completed using an encrypted person-based identifier known as the anonymised linkage field (ALF), generated by the Digital Health and Care Wales (DHCW) [33, 34]. Data preparation (extraction, cleaning, and linkage) was performed in Structured Query Language (SQL) on an IBM DB2 platform, with subsequent analyses performed in R v3.3.2 [35].

Results

Stage one – MCS

Among the consented singleton children 1,838 were assigned an ALF, with 82% of the children having a GP registration record in SAIL before age eleven years (Supplementary Figure 1). Individual and household characteristics (following imputation) are described in Table 1. 7.6% of the MCS children were considered as 'case' based on their response. Health codes (256 ICD-10 and Read codes) were obtained

after merging the first ten years of EHRs from PEDW and WLGP. Feature selection method reduced this to 13 health features (Table 2).

After merging health and socio-demographic variables, 31 main explanatory variables (13 health codes and 18 socio-demographic variables) were available for the two-way logistic model. The final 19 features with significant p values were considered to be significantly associated with the risk profile of early alcohol use (Table 3).

Stage two – whole population

In Wales, 207,114 children were born in between 1st January 1998 and 31st December 2002, and their records were obtained from WDSD. After applying exclusion criteria there were 59,231 children as the study population (Supplementary Figure 2). Of the study population, 591 (0.99%) children had at least one alcohol-related event between seven and 14 years of age (Supplementary Figure 3) who were the cases from the whole population subset. After applying case control selection, the dataset had 591cases and 11,820 non-cases, which were further split into training and test set. There were 8,688 (417 cases and 8,271 non-cases) children in the training dataset. The variables identified as significantly associated with early alcohol use using MCS data were mapped into the whole population cohort (Supplementary Table 2). Table 1 presents descriptive statistics for this population. Mothers of 6.72% of the children had an alcohol-related event reported in PEDW or WLGP while pregnant. 2.42% children lived with a household member who had alcohol-related inpatient hospital admission. The adjusted odds ratio of the features with 95% confidence interval are presented in Table 4 (also see Supplementary Figure 4).

The model was run on the test dataset. The accuracy of the model was 61.32% with a sensitivity of 58.05% and specificity of 68.48% (additional details are provided in Supplementary Tables 5, 6). Out of 174 cases, the model was able to predict 101 (58%) children who had an alcohol-related health event recorded in the healthcare system between ages seven and fourteen.

Discussion

This study has developed a two-stage data-driven framework that can create a profile of the characteristics of children who end up with an alcohol problem in adolescence. The study undertook data linkage between a longitudinal survey data (MCS) and routine EHRs in stage one to select the significant risk factors associated with early alcohol use. Stage two built the analogous risk factors using only the linked routine data and based this, a prediction model was developed. Hybridisation of these two powerful data sources (routine and survey) enabled us to create a data-driven risk profile. The risk factors were significantly associated across both MCS and whole population analyses, but effect estimates varied. Children whose health needs are supported are at lower risk of early alcohol use, evidenced by protective effect of receiving vaccinations, attending routine health examinations with their GP, and contact with health services recorded in primary and secondary care were consistent across MCS and whole population analyses. Similarly, children with health codes relating to acute upper respiratory infections may have more protective guardians willing to consult medical professionals for mild conditions. Together, this suggests that the avoidance of regular healthcare contact is an indicator that increases the risk of early alcohol use. However, the trends relating to the two codes, the child surveillance administration code and the chapter heading linked to male genitals, differed between the whole population and the MCS analysis. The code linked to male genitals showed an association with higher risk of alcohol use in MCS but was statistically inconclusive for the whole population analysis. The child surveillance administration code was associated with higher risk for the MCS cohort in contrast to the whole population which can be attributed to the differential support received by two cohorts which was not captured by the data and hence this requires further investigation. Also, the proportion of cases obtained from MCS data (stage one) were higher than those obtained from the whole population data (stage two). This can be attributed to the fact that cases from stage one were based on the self-reported alcohol consumption data whereas the stage two routine data highlighted the most severe cases caused by alcohol among the adolescents and recorded on the healthcare system.

The overall risk profile obtained from MCS and whole population analyses were broadly consistent with each other and the research literature generally both in the UK and internationally. Similar risk factors include being male [13], ever living in an urban environment where there is a greater density of alcohol outlets [36], ever living in conditions of social deprivation, living in a household with higher level of alcohol use by household members [9]. Studies from USA highlighted that early onset of alcohol use was significantly associated with parental drinking pattern and living in a lone parent household [11], child's attention deficit hyperactivity disorder (ADHD) and conduct disorder [6, 7]. The stage one MCS analysis in this study revealed that emotional difficulty and a higher level of behavioural difficulty (as assessed by parents) were associated with a reduced risk of alcohol use. However, diagnosis of clinically relevant behavioural/emotional problems was protective in the population model. Poor maternal mental health was linked with adverse outcomes, consistent with family-level risk factors that promote children's alcohol use [12, 17]. A difference was observed in regards to the effect of maternal age at birth on the risk of a child's early alcohol use. The protective effect of higher maternal age was observed for the whole population but the finding on MCS data differed and requires further investigation. Further, employment deprivation in the whole population analysis was associated with lower risk of a child's early alcohol use after adjusting for overall deprivation. This finding is similar to the existing literature [15, 37], which found that early alcohol use is more common in higher income families. This suggests that reliance on employment indicators is not sufficient to understand the socio-economic factors influencing a child's early alcohol use, the overall deprivation (also measured by education, health, access to the service, physical environment of living) plays an important role as well.

The result of this study needs to be interpreted in conjunction with a number of limitations. Firstly, mapping the MCS survey to the routine data was challenging, not all

Table 3: The explanatory variables associated with higher and lower risk of early alcohol use for the MCS children (Stage one analysis) with the adjusted Odds Ratio (OR) and 95% confidence interval (CI)

Feature	Adjusted OR (95%CI)
Child's sex	
Female	1
Male	3.06 (2.35 to 3.99)***
Mother's Socio-economic classification (SEC)	
Always Managerial or intermediate	1
Always semi-employed, self-employed, semi-routine or routine	1.30 (0.93 to 1.81)
Unknown	1.94 (1.37 to 2.74)***
Lone parent	
Never lone parent	1
Lone parent	1.68 (1.07 to 2.65)*
Ever been	1.77 (1.27 to 2.49)**
Mother alcohol use during pregnancy	
Never	1
Low (less than once a month, 1–2 times a month)	2.48 (1.83 to 3.38)***
High	5.38 (3.58 to 8.15)***
Mother alcohol use after child was born	
Never	1
Low	1.15 (0.70–1.92)
High	0.70 (0.04 to 1.24)
Guardian alcohol use	
Low	1
Moderate	1.73 (1.22 to 2.25)**
High	1.07 (0.70 to 1.64)
Variable	0.91 (0.48 to 1.70)
Living area	
Rural	1
Urban	1.61 (1.17 to 2.23)**
Ever been urban	4.54 (2.69 to 7.75)***
Poverty indicator	
Above poverty level	1
Below poverty level	0.93 (0.60 to 1.45)
Ever been below poverty level	1.33 (0.95 to 1.86)
Maternal age at child's birth	
Less than 20 years	1
20 to 24 years	1.57 (0.97 to 2.58)
25 to 29 years	3.28 (2.03 to 5.36)***
30 to 34 years	2.68 (1.64 to 4.43)***
35 years or over	0.65 (0.35 to 1.21)
Gestational age	
Not term	1
Term	9.42 (4.22 to 23.03)***
Additional household member	
No	1
Yes	0.69 (0.45 to 1.06)
Ever had	0.57 (0.39 to 0.81)**
Hyperactivity	
Always normal	1
Any mention of higher level of hyperactivity	1.84 (1.37 to 2.47)***
Conduct disorder	
Always normal	1
Any mention of higher level of CP	2.10 (1.57 to 2.82)***
Emotional difficulty	
Always normal	1
Any mention of higher level of ED	0.68 (0.48–0.97)*

(Continued).

Table 3: Continued

Feature	Adjusted OR (95%CI)
Total Difficulty Score	
Always normal	1
Any mention of higher level of TDS	0.45 (0.31 to 0.66)***
Mother longstanding illness	
No	1
Yes	1.53 (1.09 to 2.16)*
Varies	1.25 (0.96 to 1.65)
Other acute upper respiratory infections (Read code H05%)	
No	1
Yes	0.43 (0.34–0.55)***
Male genital organ diseases (Read code K2%)	
No	
Yes	2.77 (1.58–4.94)***
Child surveillance administration (Read code 919%)	
No	
Yes	1.38 (1.06 to 1.81)*
Child exam (Read code 64N%)	
No	
Yes	0.51 (0.35 to 0.75)**
Tetanus vaccination (Read code 656%)	
No	
Yes	0.60 (0.45 to 0.79)***
General examination (ICD10 code Z%)	
No	
Yes	0.73 (0.55 to 0.99)*
Disorders of penis (Read code K27%)	
No	
Yes	0.63 (0.33 to 1.19)

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

MCS variables were available in the routine data. In some instances, multiple variables had to be merged to derive summary variables. This may result in a degree of uncertainty about the information captured in the summary variables. Secondly, it was necessary to aggregate some time-varying variables into a single point estimate and, as such, the analyses are unable to capture how the recency of some events might influence results. Thirdly, due to unavailability of continuous GP records of some participants between six months and fourteen years (if the participants changed their GP and the their registered GP was not contributing to SAIL), they were removed from the whole population analysis. Similarly, the follow-up of children was not possible where they who moved out of the study area (Wales, UK), or died under age fourteen, because of which their exposure (sociodemographic and health related data) and outcome (alcohol data) data were not available. This resulted in a large reduction of the number of children in the study population. However, this did not contribute to selection bias as this happened randomly and the losses had no direct relationship with alcohol-related outcome. Fourthly, the EHRs did not include Emergency Department (ED) attendance data (but does include admissions into hospital via the ED) as there are no uniformly applicable codes for alcohol-related attendances in ED, and even when available, these are sparsely populated [38]. Lastly, in this study the model performance, measured by sensitivity and specificity, was moderate. However, even if we had a sensitivity and specificity of 90% the maximum positive predictive value, we can get is 31%, given the low prevalence of alcohol-related medical contact, as the prevalence influences the positive and negative predictive value of a model performance [39]. Machine learning approaches generally aim to achieve the best predictive models from the available data. The low positive predictive value, obtained here, suggests that the variables needed to improve model performance are not available in the data (e.g., genetic information, peer alcohol-related data).

Routine EHRs and administrative data are available to healthcare professionals and are used by policy makers and commissioners to determine how resources are best utilised to manage preventive interventions. However, the bulk of research considering early alcohol use and related outcomes has relied on self-report surveys. It has been shown that linking survey and routine data can offer new insights [40]. The results presented here are novel in that our approach generalised results from an established survey to a whole population analysis using predictive analytic techniques. This provides in-depth knowledge about the profile of the children susceptible to early alcohol use and can feasibly be used to inform population health strategies designed to reduce the

Table 4: The explanatory variables associated with higher and lower risk of early alcohol-related health outcomes for the whole population (Stage two analysis) with the adjusted Odds Ratio (OR) and 95% confidence interval (CI)

Feature	**Adjusted OR (95% CI)**
Child's Sex	
Female	1
Male	1.09 (1.02 to 1.17)**
Overall deprivation:	
Low	1
High	1.11 (0.98 to 1.25)
Borderline	1.66 (1.41 to 1.95)***
Employment deprivation:	
Low	1
High	0.84 (0.75 to 0.95)**
Borderline	0.82 (0.69 to 0.97)*
Living with single adult:	
No	1
Yes	1.45 (1.32 to 1.59)***
Ever been	1.17 (1.08 to 1.26)***
Mother's alcohol-related condition during pregnancy	
No	1
Yes	0.88 (0.77 to 1.00)*
Household member with alcohol-related condition	
No	1
Yes	2.16 (1.80 to 2.60)***
Living area	
Rural	1
Urban	0.99 (0.92 to 1.08)
Ever in urban	2.42 (2.08 to 2.81)***
Maternal age at birth	
Less than 20 years	1
20 to 24 years	0.88 (0.79 to 0.99)*
25 to 29 years	0.79 (0.71 to 0.87)***
30 to 34 years	0.68 (0.61 to 0.76)***
35 years or over	0.53 (0.46 to 0.60)***
Gestational age	
Not-term	1
Term	1.11 (0.89 to 1.40)
Child – Attention deficit hyperactive disorder (ADHD)	
No	1
Yes	3.57 (2.52 to 5.15)***
Child - Conduct disorder	
No	1
Yes	3.26 (2.14 to 5.07)***
Child – Depression/Anxiety	
No	1
Yes	0.75 (0.34 to 1.69)
Mother's any comorbidity	
No	1
Yes	1.25 (1.16 to 1.34)***
Mother's common mental health condition	
No	1
Yes	1.31 (1.23 to 1.40)***
Mother's psychosis disorder	
No	1
Yes	3.12 (2.04 to 4.90)***

(Continued).

Table 4: Continued

Feature	Adjusted OR (95% CI)
Other acute upper respiratory infections (Read code H05%)	
No	1
Yes	0.97 (0.91 to 1.04)
Male genital organ diseases (Read code K27%)	
No	1
Yes	0.90 (0.79 to 1.02)
Child surveillance administration (Read code 919%)	
No	1
Yes	0.80 (0.75 to 0.86)***
Tetanus vaccination (Read code 656%)	
No	1
Yes	0.47 (0.44 to 0.51)***
Child exam (Read code 64N%)	
No	1
Yes	0.59 (0.53 to 0.65)***
General examination (ICD10 code Z%)	
No	1
Yes	0.84 (0.78 to 0.92)***

Note: *p $<$ 0.05, **p $<$ 0.01, ***p $<$ 0.001.

prevalence of early alcohol use in children and related health outcomes.

Conclusions

The hybridisation of data of different nature, as carried out in this study, is a novel approach that combines the complementary advantages of EHRs with more personal insights from questionnaire-based cohort data. This provides a robust resource on which findings can be based and generalised to the wider population. The identified risk factors such as living with a single parent, alcohol problem in the household, social deprivation and children receiving poor support from the healthcare system indicate that involvement and support for the family is important in breaking cycles and improving children's outcomes.

Acknowledgements

This research has been carried out as part of the ADR Wales programme of work. The ADR Wales programme of work is aligned to the priority themes as identified in the Welsh Government's national strategy: Prosperity for All. ADR Wales brings together data science experts at Swansea University Medical School, staff from the Wales Institute of Social and Economic Research, Data and Methods (WISERD) at Cardiff University and specialist teams within the Welsh Government to develop new evidence which supports Prosperity for All by using the SAIL Databank at Swansea University, to link and analyse anonymised data. ADR Wales is part of the Economic and Social Research Council (part of UK Research and Innovation) funded ADR UK (grant ES/S007393/1). This work was also supported by the National Centre for Population Health and Well-Being Research (NCPHWR). The research was supported by DECIPHer, a UKCRC Public Health Research Centre of Excellence, which receives funding from the British Heart Foundation, Cancer Research UK, Medical Research Council, the Welsh Government and the Wellcome Trust (WT087640MA), under the auspices of the UK Clinical Research Collaboration. This work was supported by Health Data Research UK which receives its funding from HDR UK Ltd (NIWA1) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Welcome Trust.

The authors are grateful to the Centre for Longitudinal Studies, UCL Institute of Education and the UK Data Service. The co-operation of the participating Cohort families is also gratefully acknowledged. This work uses data provided by patients and collected by the NHS as part of their care and support. This study used anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. We would like to acknowledge all the data providers who enable SAIL to make anonymised data available for research. Compliance with ethical standards.

Funding

This work was supported by funds from the Economic and Social Research Council, the Medical Research Council and Alcohol Research UK to the ELAStiC Project (ES/L015471/1).

The study funders had no involvement in the study design; the collection, analysis, and interpretation of data; the writing

of the report; and the decision to submit the paper for publication.

Dedication

This work was designed with Professor Damon Berridge. Damon passed away April 12th, 2019, and is greatly missed by us all.

Contributorship statement

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Amrita Bandyopadhyay and Sinead Brophy. The first draft of the manuscript was written by Amrita Bandyopadhyay, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conceptualization: Sinead Brophy and Amrita Bandyopadhyay; Methodology: Amrita Bandyopadhyay, Damon Berridge, and Sinead Brophy; Formal analysis and investigation: Amrita Bandyopadhyay Writing - original draft preparation: Amrita Bandyopadhyay; Writing - review and editing: Simon Moore, Sinead Brophy, Ashley Akbari, Joanne Demmler, Shantini Paranjothy, Jonathan Kennedy and Ronan A Lyons; Funding acquisition: Simon Moore, Shantini Paranjothy and Ronan A Lyons; Resources: Ashley Akbari; Supervision: Sinead Brophy and Simon Moore.

Conflict of interest

The authors declare that they have no conflict of interest.

Ethics statement

Ethics approval for the fourth survey of the Millennium Cohort Study was received from the Northern and Yorkshire Research Ethics Committee (07/MRE03/32). This study was approved by the SAIL Databank independent Information Governance Review Panel (IGRP) (project number 0336).

References

1. Hingson RW, Heeren T, Winter MR. Age at Drinking Onset and Alcohol Dependence: Age at Onset, Duration, and Severity. *Arch Pediatr Adolesc Med*. 2006; 160(7):739–746. https://doi.org/10.1001/archpedi.160.7.739

2. Bi J, Sun J, Wu Y, Tennen H, Armeli S. A Machine Learning Approach to College Drinking Prediction and Risk Factor Identification. *ACM Trans Intell Syst Technol*. 2013;4(4):72:1–72:24. https://doi.org/10.1145/2508037.2508053

3. Hingson RW, Zha W, Weitzman ER. Magnitude of and Trends in Alcohol-Related Mortality and Morbidity Among U.S. College Students Ages 18–24, 1998–2005. *J Stud Alcohol Drugs Suppl*. 2009;(16):12–20. Accessed August 2, 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2701090/

4. National Institute on Alcohol Abuse and Alcoholism. Underage Drinking: A Major Public Health Challenge – Alcohol Alert No. 59. Published April 2003. Accessed August 2, 2018. https://pubs.niaaa.nih.gov/publications/aa59.htm

5. Office of the Surgeon General (US), National Institute on Alcohol Abuse and Alcoholism (US), Substance Abuse and Mental Health Services Administration (US). *The Surgeon General's Call to Action To Prevent and Reduce Underage Drinking*. Office of the Surgeon General (US); 2007. Accessed April 3, 2019. http://www.ncbi.nlm.nih.gov/books/NBK44360/

6. Molina BS, Pelham WE. Childhood predictors of adolescent substance use in a longitudinal study of children with ADHD. *J Abnorm Psychol*. 2003;112(3): 497–507.

7. Sibley MH, Pelham WE, Molina BSG, et al. The role of early childhood ADHD and subsequent CD in the initiation and escalation of adolescent cigarette, alcohol, and marijuana use. *J Abnorm Psychol*. 2014;123(2):362–374. https://doi.org/10.1037/a0036585

8. Kelly Y, Goisis A, Sacker A, Cable N, Watt RG, Britton A. What influences 11-year-olds to drink? Findings from the Millennium Cohort Study. *BMC Public Health*. 2016;16(1):169. https://doi.org/10.1186/s12889-016-2847-x

9. Mahedy L, MacArthur GJ, Hammerton G, et al. The effect of parental drinking on alcohol use in young adults: the mediating role of parental monitoring and peer deviance. *Addiction*. 2018;113(11):2041–2050. https://doi.org/10.1111/add.14280

10. Simantov E, Schoen C, Klein JD. Health-Compromising Behaviors: Why Do Adolescents Smoke or Drink?: Identifying Underlying Risk and Protective Factors. *Arch Pediatr Adolesc Med*. 2000;154(10):1025–1033. https://doi.org/10.1001/archpedi.154.10.1025

11. Donovan JE, Molina BSG. Childhood Risk Factors for Early-Onset Drinking*. *J Stud Alcohol Drugs*. 2011;72(5):741–751. https://doi.org/10.15288/jsad.2011.72.741

12. Dube SR, Miller JW, Brown DW, et al. Adverse childhood experiences and the association with ever using alcohol and initiating alcohol use during adolescence. *J Adolesc Health*. 2006;38(4):444.e1–444.e10. https://doi.org/10.1016/j.jadohealth.2005.06.006

13. Kelly Y, Britton A, Cable N, Sacker A, Watt RG. Drunkenness and heavy drinking among 11year olds - Findings from the UK Millennium Cohort Study. *Prev Med*. 2016;90:139–142. https://doi.org/10.1016/j.ypmed.2016.07.010

14. Marshall EJ. Adolescent Alcohol Use: Risks and Consequences. *Alcohol Alcohol*. 2014;49(2):160–164. https://doi.org/10.1093/alcalc/agt180

15. Melotti R, Heron J, Hickman M, Macleod J, Araya R, Lewis G. Adolescent Alcohol and Tobacco Use and Early Socioeconomic Position: The ALSPAC Birth Cohort. *Pediatrics*. 2011;127(4):e948–e955. https://doi.org/10.1542/peds.2009-3450

16. Mersky JP, Topitzes J, Reynolds AJ. Impacts of adverse childhood experiences on health, mental health, and substance use in early adulthood: A cohort study of an urban, minority sample in the U.S. *Child Abuse Negl*. 2013;37(11):917–925. https://doi.org/10.1016/j.chiabu.2013.07.011

17. Paranjothy S, Evans A, Bandyopadhyay A, et al. Risk of emergency hospital admission in children associated with mental disorders and alcohol misuse in the household: an electronic birth cohort study. *Lancet Public Health*. 2018;3(6):e279–e288. https://doi.org/10.1016/S2468-2667(18)30069-0

18. Huber M, Knottnerus JA, Green L, et al. How should we define health? *BMJ*. 2011;343:d4163. https://doi.org/10.1136/bmj.d4163

19. Kell DB, Oliver SG. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays*. 2004;26(1):99–105. https://doi.org/10.1002/bies.10385

20. Saracci R. Epidemiology in wonderland: Big Data and precision medicine. *Eur J Epidemiol*. 2018;33(3):245–257. https://doi.org/10.1007/s10654-018-0385-9

21. Sedgwick P. Questionnaire surveys: sources of bias. *BMJ*. 2013;347:f5265. https://doi.org/10.1136/bmj.f5265

22. Sivarajah U, Kamal MM, Irani Z, Weerakkody V. Critical analysis of Big Data challenges and analytical methods. *J Bus Res*. 2017;70:263–286. https://doi.org/10.1016/j.jbusres.2016.08.001

23. Connelly R, Platt L. Cohort Profile: UK Millennium Cohort Study (MCS). *Int J Epidemiol*. 2014;43(6):1719–1725. https://doi.org/10.1093/ije/dyu001

24. Goodman A, Goodman R. Strengths and Difficulties Questionnaire as a Dimensional Measure of Child Mental Health. *J Am Acad Child Adolesc Psychiatry*. 2009; 48(4):400–403. https://doi.org/10.1097/CHI.0b013e3181985068

25. Hyatt M, Rodgers SE, Paranjothy S, Fone D, Lyons RA. The wales electronic cohort for children (WECC) study. *Arch Dis Child - Fetal Neonatal Ed*. 2011;96(Suppl 1):Fa18–Fa18. https://doi.org/10.1136/archdischild.2011.300164.6

26. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw*. Published online 2010:1–68.

27. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011;30(4):377–399. https://doi.org/10.1002/sim.4067

28. Cantú-Paz E, Newsam S, Kamath C. Feature Selection in Scientific Applications. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '04. ACM; 2004:788–793. https://doi.org/10.1145/1014052.1016915

29. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media; 2003.

30. Rodgers SE, Demmler JC, Dsilva R, Lyons RA. Protecting health data privacy while using residence-based environment and demographic data. *Health Place*. 2012;18(2):209–217. https://doi.org/10.1016/j.healthplace.2011.09.006

31. Trefan L, Akbari A, Paranjothy S, et al. Electronic Longitudinal Alcohol Study in Communities (ELAStiC) Wales – protocol for platform development. *Int J Popul Data Sci*. 2019;4(1). https://doi.org/10.23889/ijpds.v4i1.581

32. Rose S, van der Laan MJ. Why Match? Investigating Matched Case-Control Study Designs with Causal Effect Estimation. *Int J Biostat*. 2009;5(1). https://doi.org/10.2202/1557-4679.1127

33. Ford DV, Jones KH, Verplancke JP, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res*. 2009;9:157. https://doi.org/10.1186/1472-6963-9-157

34. Lyons RA, Jones KH, John G, et al. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak*. 2009;9(1):3. https://doi.org/10.1186/1472-6947-9-3

35. R Core Team. R: A Language and Environment for Statistical Computing. Published 2018. Accessed November 22, 2018. https://doi.org/10.1186/1472-6947-9-3

36. Gartner A, Farewell DM, Morgan J, et al. Association between alcohol outlet density and alcohol-related mortality in Wales: an e-cohort study. *The Lancet*. 2017;390:S14. https://doi.org/10.1016/S0140-6736(17)32949-5

37. Moore SC, Orpen B, Smith J, et al. Alcohol affordability: implications for alcohol price policies. A cross-sectional analysis in middle and older adults from UK Biobank. *J Public Health*. 2021;(fdab095). https://doi.org/10.1093/pubmed/fdab095

38. Fone D, Dunstan F, White J, et al. Change in alcohol outlet density and alcohol-related harm to population health (CHALICE). *BMC Public Health*. 2012;12(1):428. https://doi.org/10.1186/1471-2458-12-428

39. Brenner H, Gefeller O. Variation of Sensitivity, Specificity, Likelihood Ratios and Predictive Values with Disease Prevalence. *Stat Med*. 1997;16(9):981–991. https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<981::AID-SIM510>3.0.CO;2-N

40. Gray L, Batty GD, Craig P, et al. Cohort Profile: The Scottish Health Surveys Cohort: linkage of study participants to routinely collected records for mortality, hospital discharge, cancer and offspring birth characteristics in three nationwide studies. *Int J Epidemiol*. 2010;39(2):345–350. https://doi.org/10.1093/ije/dyp155

Abbreviations

ALF: Anonymised linkage field #
ED: Emergency Department
EHR: Electronic health record
LSOA: Lower super output area
MCS: Millennium Cohort Study
NWIS: National Health Service Wales Informatics Service
PEDW: Patient Episode Database for Wales
PMM: predictive mean matching
RALF: Residential anonymised linking field
SAIL: Secure Anonymised Information Linkage
SDQ: Strength and Difficulty Questionnaire
SEC: socio-economic classification
SQL: Structured Query Language
WDSD: Welsh Demographic Service Dataset
WECC: Wales Electronic Cohort for Children
WLGP: Welsh Longitudinal General Practice
WIMD: Welsh Index of Multiple Deprivation

Supplementary Appendices

Supplementary table 1: MCS alcohol-related questions and criteria for inclusion in the case group

Questions	Criteria
How many times have you had an alcoholic drink in the last 12 months?	3–5 times or more
How many times have you had an alcoholic drink in the last four weeks?	1–2 times or more
Have you ever drunk enough to feel drunk?	Yes
Have you ever had five or more alcoholic drinks at a time? A drink is half a pint of lager, beer or cider, one alcopop, a small glass of wine, or a measure of spirits.	Yes
How many times have you had five or more alcoholic drinks at a time?	Once or more

Supplementary table 2: MCS to Whole Population explanatory variables mapping

MCS Predictor	Whole Population Analogue	Source	Time Varying	Code	Method
Gender	Gender	WDSD	No	1 = male 0 = female	
Lone parent	Living with single adult	WDSD	Yes		
Additional household member	Living with single adult	WDSD	Yes	0 = Never with a single adult 1 = Always with a single adult 2 = Ever been with single adult	1. Using RALF, number of people sharing same house with child at the above mentioned 4 time points were derived 2. Based on household members' age at the 4 time points, the number of adults staying with child was determined 3. A binary variable was created based on the number adults at the household at 4 time points 4. A categorical summary variable was created to identify the overall status of the concept variable
Mother's SEC	Employment deprivation	WIMD reference data from Welsh Government	Yes	0 = Always in least deprived group 1 = Always in most deprived group 2 = Ever belong to most deprived group	1. Welsh Index of Multiple Deprivation (WIMD) quintile scale on employment and overall deprivation at each time point for each RALF was achieved. 2. WIMD quintile scale between 1 and 5 (from most to least deprivation). 3. The study combined the scale 1 and 2 to indicate the most deprived group and the rest 3 scales were classified as least deprived group 4. A categorical summary variable was created to identify the overall status of the concept variable

(Continued).

Supplementary table 2: Continued

MCS Predictor	Whole Population Analogue	Source	Time Varying	Code	Method
Mother alcohol use during pregnancy	Mother's alcohol-related condition during pregnancy	WECC, WLGP, PEDW	No	1 = Yes 0 = No	1. From WECC the maternal ALF was obtained 2. Based on gestational age and the week of birth the pregnancy period was calculated 3. If the mother had an alcohol-related code recorded in WLGP or PEDW during the pregnancy period then a binary flag variable was created
Guardian alcohol use	Household member with alcohol-related hospital admission record	WDSD, PEDW	Yes	0 = Never lived with someone who had an alcohol hospital admission 1 = Ever lived with someone who had an alcohol hospital admission	1. Using RALF, any household member had an alcohol-related event recorded in WLGP or PEDW between birth to < nine months, nine months to < three years, three years to < five years and five years to < seven years -was identified 2. A categorical summary variable was created
Living area	Living area	WDSD and Rural Urban indicator reference data from Welsh Government	Yes	0 = Always lived in rural area 1 = Always lived in urban area 2 = Ever lived in urban area	1. Each RALF is always within a Lower super Output Area (LSOA) code. 2. Each LSOA code is further categorised using the rural urban indicators into urban, village and town. 3. In this study village and town are grouped together and classified as rural. 4. A categorical summary variable was created
Maternal age at birth	Maternal age at birth	WECC	No	Less than 20 years 20 to 24 years 25 to 29 years 30 to 34 years 35 years or over	
Gestational age	Gestational age	WECC	No	1 = not term 0 = term	

(Continued).

Supplementary table 2: Continued

MCS Predictor	Whole Population Analogue	Source	Time Varying	Code	Method
Mother longstanding illness	Mother's any comorbidity Mother's psychosis disorder	WLGP, PEDW	No	1 = yes	Any longstanding health condition, common mental health condition and psychosis disorder between their birth and the seven years of their child's age
	Mother's common mental health condition			0 = no	
Conduct disorder	Conduct disorder (CD)	WLGP	No	1 = yes	CD diagnosis/treatment by GP between birth and age seven
				0 = no	
Hyperactivity	Attention Deficit Hyperactivity disorder (ADHD)	WLGP	No	1 = yes	ADHD diagnosis/treatment by GP between birth and age seven
				0 = no	
Emotional difficulty Total difficulty score	Other mental health condition	WLGP, PEDW	No	1 = yes	Any mental health condition (apart from ADHD and CD codes) reported in GP
				0 = no	Any mental health condition related hospital admission between birth and age seven
Health codes: 5 Read codes and 1 ICD10 codes	Health codes: 5 Read codes and 1 ICD10 codes	Read codes from WLGP and ICD10 codes from PEDW	No		Individual code recorded in WLGP and PEDW between birth and age 7

Supplementary table 3: Alcohol-related ICD10 codes

ICD10 Code	Description
E244	Alcohol-induced pseudo-Cushing's syndrome
E512	Wernicke's encephalopathy
F100	Mental and behavioural disorders due to use of alcohol
F101	Mental and behavioural disorders due to use of alcohol
F102	Mental and behavioural disorders due to use of alcohol
F103	Mental and behavioural disorders due to use of alcohol
F104	Mental and behavioural disorders due to use of alcohol
F105	Mental and behavioural disorders due to use of alcohol
F106	Mental and behavioural disorders due to use of alcohol
F107	Mental and behavioural disorders due to use of alcohol
F108	Mental and behavioural disorders due to use of alcohol
F109	Mental and behavioural disorders due to use of alcohol
G312	Degeneration of nervous system due to alcohol
G405	Special epileptic syndromes
G621	Alcoholic polyneuropathy
G721	Alcoholic myopathy
I426	Alcoholic cardiomyopathy
K292	Alcoholic gastritis
K700	Alcoholic fatty liver
K701	Alcoholic hepatitis
K702	Alcoholic fibrosis and sclerosis of liver
K703	Alcoholic cirrhosis of liver
K704	Alcoholic hepatic failure
K709	Alcoholic liver disease, unspecified
K852	Alcohol-induced acute pancreatitis
K860	Alcohol-induced chronic pancreatitis
O354	Maternal care for (suspected) damage to fetus from alcohol
Q860	Fetal alcohol syndrome (dysmorphic)
R780	Finding of alcohol in blood
T510	Toxic effect: Ethanol
X450–X459	Accidental poisoning by and exposure to alcohol
X650–X659	Intentional self-poisoning by and exposure to alcohol
Y150	Poisoning by and exposure to alcohol, undetermined intent
Y152	Poisoning by and exposure to alcohol, undetermined intent
Y154	Poisoning by and exposure to alcohol, undetermined intent
Y158	Poisoning by and exposure to alcohol, undetermined intent
Y159	Poisoning by and exposure to alcohol, undetermined intent
Y900	Blood alcohol level of less than 20 mg/100 ml
Y901	Blood alcohol level of 20–39 mg/100 ml
Y902	Blood alcohol level of 40–59 mg/100 ml
Y903	Blood alcohol level of 60–79 mg/100 ml
Y904	Blood alcohol level of 80–99 mg/100 ml
Y905	Blood alcohol level of 100–119 mg/100 ml
Y906	Blood alcohol level of 120–199 mg/100 ml
Y907	Blood alcohol level of 200–239 mg/100 ml
Y908	Blood alcohol level of 240 mg/100 ml or more
Y909	Presence of alcohol in blood, level not specified
Y910	Mild alcohol intoxication
Y911	Moderate alcohol intoxication
Y912	Severe alcohol intoxication
Y913	Very severe alcohol intoxication
Y919	Alcohol involvement, not otherwise specified
Z502	Alcohol rehabilitation
Z714	Alcohol abuse counselling and surveillance
Z721	Alcohol use

Supplementary table 4: Alcohol-related read codes

Read Code	Description
136..	Alcohol consumption
1362	Trivial drinker – <1 u/day
1363	Light drinker – 1–2 u/day
1364	Moderate drinker – 3–6 u/day
1365	Heavy drinker – 7–9 u/day
1366	Very heavy drinker – >9 u/day
1368	Alcohol consumption unknown
1369	Suspect alcohol abuse – denied
136F.	Spirit drinker
136G.	Beer drinker
136H.	Drinks beer and spirits
136I.	Drinks wine
136J.	Social drinker
136K.	Alcohol intake above recommended sensible limits
136L.	Alcohol intake within recommended sensible limits
136N.	Light drinker
136O.	Moderate drinker
136P.	Heavy drinker
136Q.	Very heavy drinker
136R.	Binge drinker
136S.	Hazardous alcohol use
136T.	Harmful alcohol use
136V.	Alcohol units per week
136W.	Alcohol misuse
136X.	Alcohol units consumed on heaviest drinking day
136Y.	Drinks in morning to get rid of hangover
136Z.	Alcohol consumption NOS
136a.	Increasing risk drinking
136b.	Feels should cut down drinking
136c.	Higher risk drinking
136d.	Lower risk drinking
136e.	Declines to state current alcohol consumption
13Y8.	Alcoholics anonymous
13ZY.	Disqualified from driving due to excess alcohol
1462	H/O: alcoholism
1B1c.	Alcohol induced hallucinations
1F9D.	Replaces meals with drinks
2126C	Alcohol dependence resolved
2577	O/E – breath – alcohol smell
388u.	Fast alcohol screening test
38D2.	Single alcohol screening questionnaire
38D3.	Alcohol use disorders identification test
38D4.	Alcohol use disorder identification test consumption questionnaire
38D5.	Alcohol use disorder identification test Piccinelli consumption questionnaire
38Df.	Five-shot questionnaire on heavy drinking
38Dz.	Severity of alcohol dependence questionnaire
38P03	Health of the Nation Outcome Scale for Children and Adolescents item 4 – alcohol, substance/solvent misuse
38QA.	CIWA-Ar - Clinical Institute Withdrawal Assessment for Alcohol scale, revised
38QE.	Addiction Research Foundation Clinical Institute Withdrawal Assessment for Alcohol
44X3.	Blood ethanol level
66e..	Alcohol disorder monitoring
66e0.	Alcohol abuse monitoring
6792	Health ed. – alcohol
67A5.	Pregnancy alcohol advice
67H0.	Lifestyle advice regarding alcohol

(Continued).

Supplementary table 4: Continued

Read Code	Description
67K6.	Cycle of change stage, alcohol
6892	Alcohol consumption screen
68S..	Alcohol consumption screen
7P221	Delivery of rehabilitation for alcohol addiction
8BA8.	Alcohol detoxification
8BAs.	Alcohol relapse prevention
8BAu.	Alcohol harm reduction programme
8CAM.	Patient advised about alcohol
8CAM0	Advised to abstain from alcohol consumption
8CAv.	Advised to contact primary care alcohol worker
8CE1.	Alcohol leaflet given
8CdK.	Specialist alcohol treatment service signposted
8G32.	Aversion therapy – alcoholism
8H35.	Admitted to alcohol detoxification centre
8H7p.	Referral to community alcohol team
8HHe.	Referral to community drug and alcohol team
8HkG.	Referral to specialist alcohol treatment service
8HkJ.	Referral to alcohol brief intervention service
8IA7.	Alcohol consumption screening test declined
8IAF.	Brief intervention for excessive alcohol consumption declined
8IAJ.	Declined referral to specialist alcohol treatment service
8IAt.	Extended intervention for excessive alcohol consumption declined
8IEA.	Referral to community alcohol team declined
8IH4.	Alcohol Use Disorders Identification Test declined
8W2..	Referral to mental health services deferred until alcohol misuse resolved
9EQ..	HO/RTS-police:venesect alc
9EVD.	Hospital alcohol liaison team report received
9NJz.	In-house alcohol detoxification
9NN2.	Under care of community alcohol team
9NgzH	Withdrawn from alcohol detoxification programme
9NzA.	Hospital attendance related to personal alcohol consumption
9k1..	Alcohol misuse – enhanced services administration
9k11.	Alcohol consumption counselling
9k12.	Alcohol misuse – enhanced service completed
9k13.	Alcohol questionnaire completed
9k14.	Alcohol counselling by other agencies
9k15.	Alcohol screen – alcohol use disorder identification test completed
9k16.	Alcohol screen – fast alcohol screening test completed
9k17.	Alcohol screen – alcohol use disorder identification test consumption questions completed
9k18.	Alcohol screen – alcohol use disorder identification test Piccinelli consumption questions completed
9k19.	Alcohol assessment declined – enhanced services administration
9k1A.	Brief intervention for excessive alcohol consumption completed
9k1B.	Extended intervention for excessive alcohol consumption completed
C1505	Alcohol-induced pseudo-Cushing's syndrome
E01..	Alcoholic psychoses
E010.	Alcohol withdrawal delirium
E011.	Alcohol amnestic syndrome
E0110	Korsakov's alcoholic psychosis
E0111	Korsakov's alcoholic psychosis with peripheral neuritis
E011z	Alcohol amnestic syndrome NOS
E012.	Other alcoholic dementia
E0120	Chronic alcoholic brain syndrome
E013.	Alcohol withdrawal hallucinosis
E014.	Pathological alcohol intoxication
E015.	Alcoholic paranoia

(Continued).

Supplementary table 4: Continued

Read Code	Description
E01y.	Other alcoholic psychosis
E01y0	Alcohol withdrawal syndrome
E01yz	Other alcoholic psychosis NOS
E01z.	Alcoholic psychosis NOS
E23..	Alcohol dependence syndrome
E230.	Acute alcoholic intoxication in alcoholism
E2300	Acute alcoholic intoxication, unspecified, in alcoholism
E2301	Continuous acute alcoholic intoxication in alcoholism
E2302	Episodic acute alcoholic intoxication in alcoholism
E2303	Acute alcoholic intoxication in remission, in alcoholism
E230z	Acute alcoholic intoxication in alcoholism NOS
E231.	Chronic alcoholism
E2310	Unspecified chronic alcoholism
E2311	Continuous chronic alcoholism
E2312	Episodic chronic alcoholism
E2313	Chronic alcoholism in remission
E231z	Chronic alcoholism NOS
E23z.	Alcohol dependence syndrome NOS
E250.	Nondependent alcohol abuse
E2500	Nondependent alcohol abuse, unspecified
E2501	Nondependent alcohol abuse, continuous
E2502	Nondependent alcohol abuse, episodic
E2503	Nondependent alcohol abuse in remission
E250z	Nondependent alcohol abuse NOS
Eu10.	[X]Mental and behavioural disorders due to use of alcohol
Eu100	[X]Mental and behavioural disorders due to use of alcohol: acute intoxication
Eu101	[X]Mental and behavioural disorders due to use of alcohol: harmful use
Eu102	[X]Mental and behavioural disorders due to use of alcohol: dependence syndrome
Eu103	[X]Mental and behavioural disorders due to use of alcohol: withdrawal state
Eu104	[X]Mental and behavioural disorders due to use of alcohol: withdrawal state with delirium
Eu105	[X]Mental and behavioural disorders due to use of alcohol: psychotic disorder
Eu106	[X]Mental and behavioural disorders due to use of alcohol: amnesic syndrome
Eu107	[X]Mental and behavioural disorders due to use of alcohol: residual and late-onset psychotic disorder
Eu108	[X]Alcohol withdrawal-induced seizure
Eu10y	[X]Mental and behavioural disorders due to use of alcohol: other mental and behavioural disorders
Eu10z	[X]Mental and behavioural disorders due to use of alcohol: unspecified mental and behavioural disorder
F11x0	Cerebral degeneration due to alcoholism
F1440	Cerebellar ataxia due to alcoholism
F25B.	Alcohol-induced epilepsy
F375.	Alcoholic polyneuropathy
F3941	Alcoholic myopathy
G555.	Alcoholic cardiomyopathy
G8523	Oesophageal varices in alcoholic cirrhosis of the liver
J153.	Alcoholic gastritis
J610.	Alcoholic fatty liver
J611.	Acute alcoholic hepatitis
J612.	Alcoholic cirrhosis of liver
J6120	Alcoholic fibrosis and sclerosis of liver
J613.	Alcoholic liver damage unspecified
J6130	Alcoholic hepatic failure
J617.	Alcoholic hepatitis
J6170	Chronic alcoholic hepatitis
J6708	Alcohol-induced acute pancreatitis
J6710	Alcohol-induced chronic pancreatitis

(Continued).

Supplementary table 4: Continued

Read Code	Description
L2553	Maternal care for (suspected) damage to fetus from alcohol
PK80.	Fetal alcohol syndrome
PK83.	Fetus and newborn affected by maternal use of alcohol
Q0071	Fetus or neonate affected by placental or breast transfer of alcohol
R103.	[D]Alcohol blood level excessive
SLH3.	Alcohol deterrent poisoning
SM0..	Alcohol causing toxic effect
SM00.	Ethyl alcohol causing toxic effect
SM000	Ethanol causing toxic effect
SM002	Grain alcohol causing toxic effect
SM00z	Ethyl alcohol causing toxic effect NOS
SM0z.	Alcohol causing toxic effect NOS
T90..	Accidental poisoning by alcohol, NEC
T900.	Accidental poisoning by alcoholic beverages
T901.	Accidental poisoning by other ethyl alcohol and its products
T9012	Accidental poisoning by grain alcohol NOS
T901z	Accidental poisoning by ethyl alcohol NOS
T90z.	Accidental poisoning by alcohol NOS
TJH3.	Adverse reaction to alcohol deterrents
U1A9.	[X]Accidental poisoning by and exposure to alcohol
U1A90	[X]Accidental poisoning by and exposure to alcohol, occurrence at home
U1A91	[X]Accidental poisoning by and exposure to alcohol, occurrence in residential institution
U1A92	[X]Accidental poisoning by and exposure to alcohol, occurrence at school, other institution and public administrative area
U1A93	[X]Accidental poisoning by and exposure to alcohol, occurrence at sports and athletics area
U1A94	[X]Accidental poisoning by and exposure to alcohol, occurrence on street and highway
U1A95	[X]Accidental poisoning by and exposure to alcohol, occurrence at trade and service area
U1A96	[X]Accidental poisoning by and exposure to alcohol, occurrence at industrial and construction area
U1A97	[X]Accidental poisoning by and exposure to alcohol, occurrence on farm
U1A9y	[X]Accidental poisoning by and exposure to alcohol, occurrence at other specified place
U1A9z	[X]Accidental poisoning by and exposure to alcohol, occurrence at unspecified place
U209.	[X]Intentional self poisoning by and exposure to alcohol
U2090	[X]Intentional self poisoning by and exposure to alcohol, occurrence at home
U2091	[X]Intentional self poisoning by and exposure to alcohol, occurrence in residential institution
U2092	[X]Intentional self poisoning by and exposure to alcohol, occurrence at school, other institution and public administrative area
U2093	[X]Intentional self poisoning by and exposure to alcohol, occurrence at sports and athletics area
U2094	[X]Intentional self poisoning by and exposure to alcohol, occurrence on street and highway
U2095	[X]Intentional self poisoning by and exposure to alcohol, occurrence at trade and service area
U2096	[X]Intentional self poisoning by and exposure to alcohol, occurrence at industrial and construction area
U2097	[X]Intentional self poisoning by and exposure to alcohol, occurrence on farm
U209y	[X]Intentional self poisoning by and exposure to alcohol, occurrence at other specified place
U209z	[X]Intentional self poisoning by and exposure to alcohol, occurrence at unspecified place
U4097	[X]Poisoning by and exposure to alcohol, occurrence on farm, undetermined intent
U60H3	[X]Alcohol deterrents causing adverse effects in therapeutic use
U8...	[X]Supplementary factors related to causes of morbidity and mortality classified elsewhere
U80..	[X]Evidence of alcohol involvement determined by blood alcohol level
U800.	[X]Evidence of alcohol involvement determined by blood alcohol level of less than 20 mg/100 ml
U801.	[X]Evidence of alcohol involvement determined by blood alcohol level of 20–39 mg/100 ml
U802.	[X]Evidence of alcohol involvement determined by blood alcohol level of 40–59 mg/100 ml
U803.	[X]Evidence of alcohol involvement determined by blood alcohol level of 60–79 mg/100 ml
U804.	[X]Evidence of alcohol involvement determined by blood alcohol level of 80–99vmg/100 ml
U805.	[X]Evidence of alcohol involvement determined by blood alcohol level of 100–119 mg/100 ml
U806.	[X]Evidence of alcohol involvement determined by blood alcohol level of 120–199 mg/100 ml

(Continued).

Supplementary table 4: Continued

Read Code	Description
U807.	[X]Evidence of alcohol involvement determined by blood alcohol level of 200–239 mg/100 ml
U808.	[X]Evidence of alcohol involvement determined by blood alcohol level of 240 mg/100 ml or more
U80z.	[X]Evidence of alcohol involvement determined by presence of alcohol in blood, level not specified
U81..	[X]Evidence of alcohol involvement determined by level of intoxication
U810.	[X]Evidence of alcohol involvement determined by level of intoxication, mild alcohol intoxication
U811.	[X]Evidence of alcohol involvement determined by level of intoxication, moderate alcohol intoxication
U812.	[X]Evidence of alcohol involvement determined by level of intoxication, severe alcohol intoxication
U813.	[X]Evidence of alcohol involvement determined by level of intoxication, very severe alcohol intoxication
U814.	[X]Evidence of alcohol involvement determined by level of intoxication, alcohol involvement, not otherwise specified
ZV113	[V]Personal history of alcoholism
ZV4KC	[V] Alcohol use
ZV57A	[V]Alcohol rehabilitation
ZV6D6	[V]Alcohol abuse counselling and surveillance
ZV704	[V]Medicolegal examination
ZV70L	[V]Blood-alcohol and blood-drug test
ZV791	[V]Screening for alcoholism
du11.	DISULFIRAM 200 mg tablets
du12.	ANTABUSE 200 mg tablets

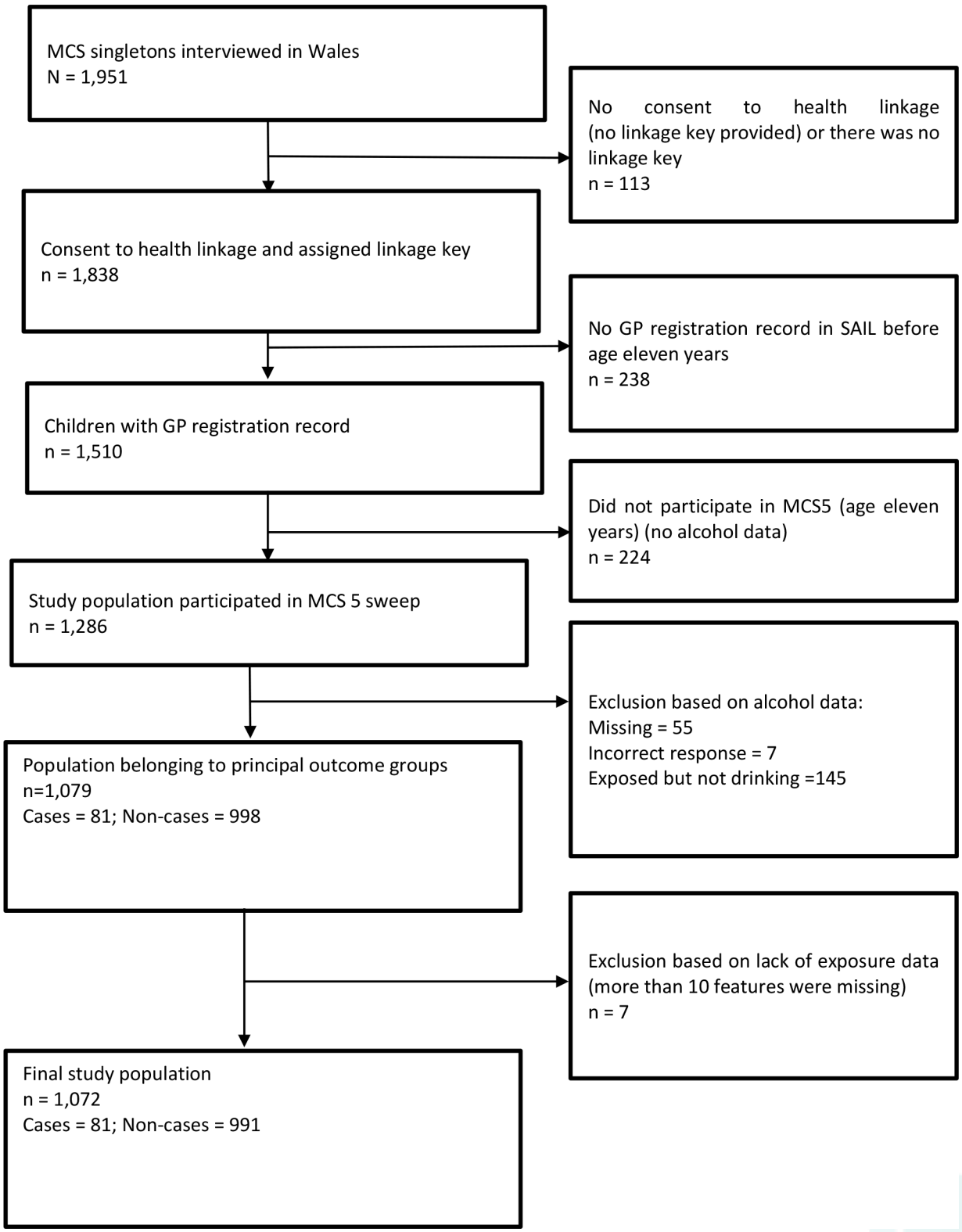
Supplementary table 5: The contingency table for the whole population analysis

	Actual negative	Actual positive	Total
Predicted negative	2,182 (true negative [TN])	73 (false negative [FN])	2,255
Predicted positive	1,367 (false positive [FP])	101 (true positive [TP])	1,468
Total	3,549	174	3,723

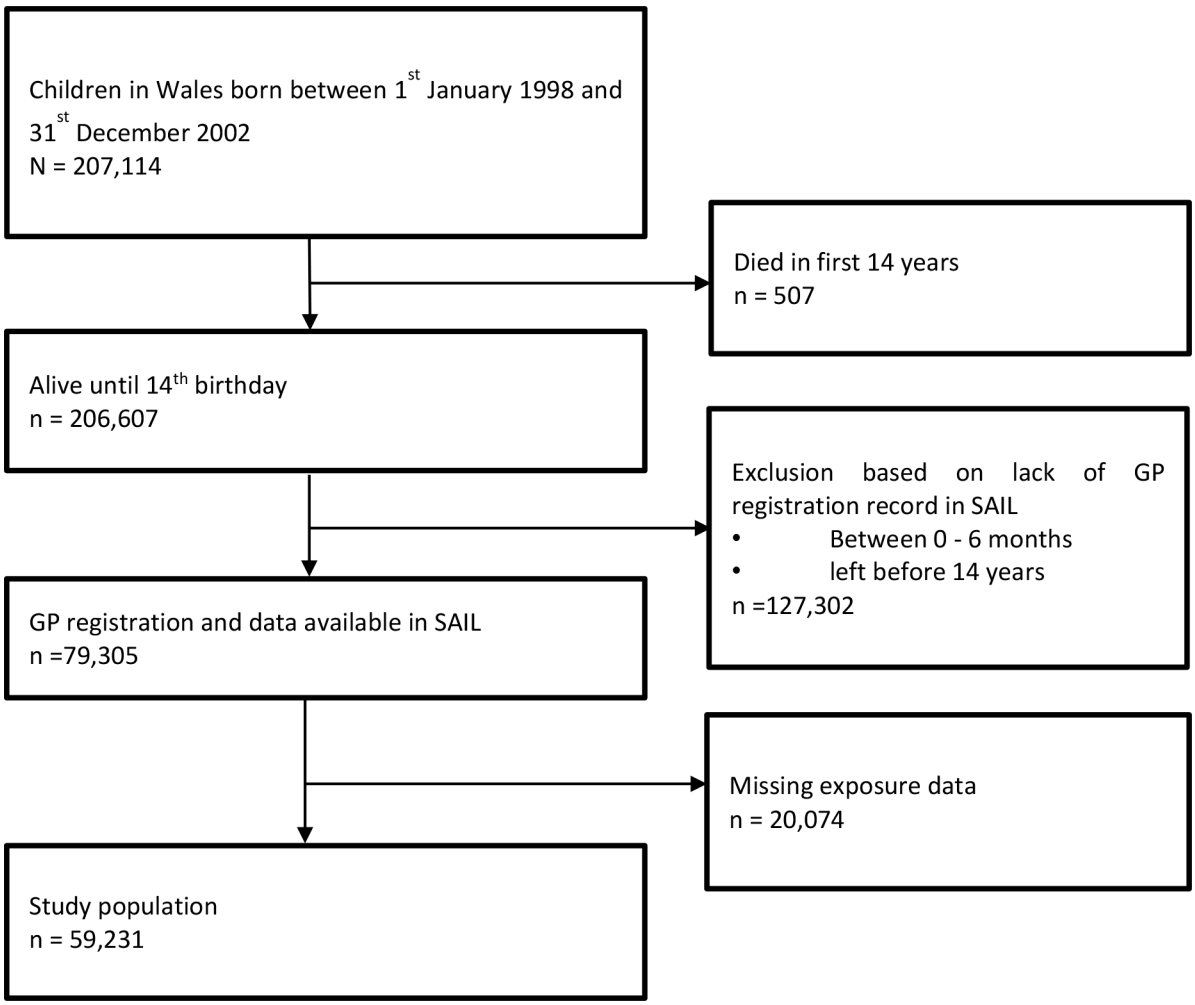
Supplementary table 6: Model prediction results

Measurement	Formula	Value
Accuracy	TP + TN / TP + TN + FP + FN	61.32
Sensitivity	TP / TP + FN	58.05
Specificity	TN / TN + FP	61.48
Positive predictive value	TP / TP + FP	6.88
Negative predictive value	TN / TN + FN	96.76

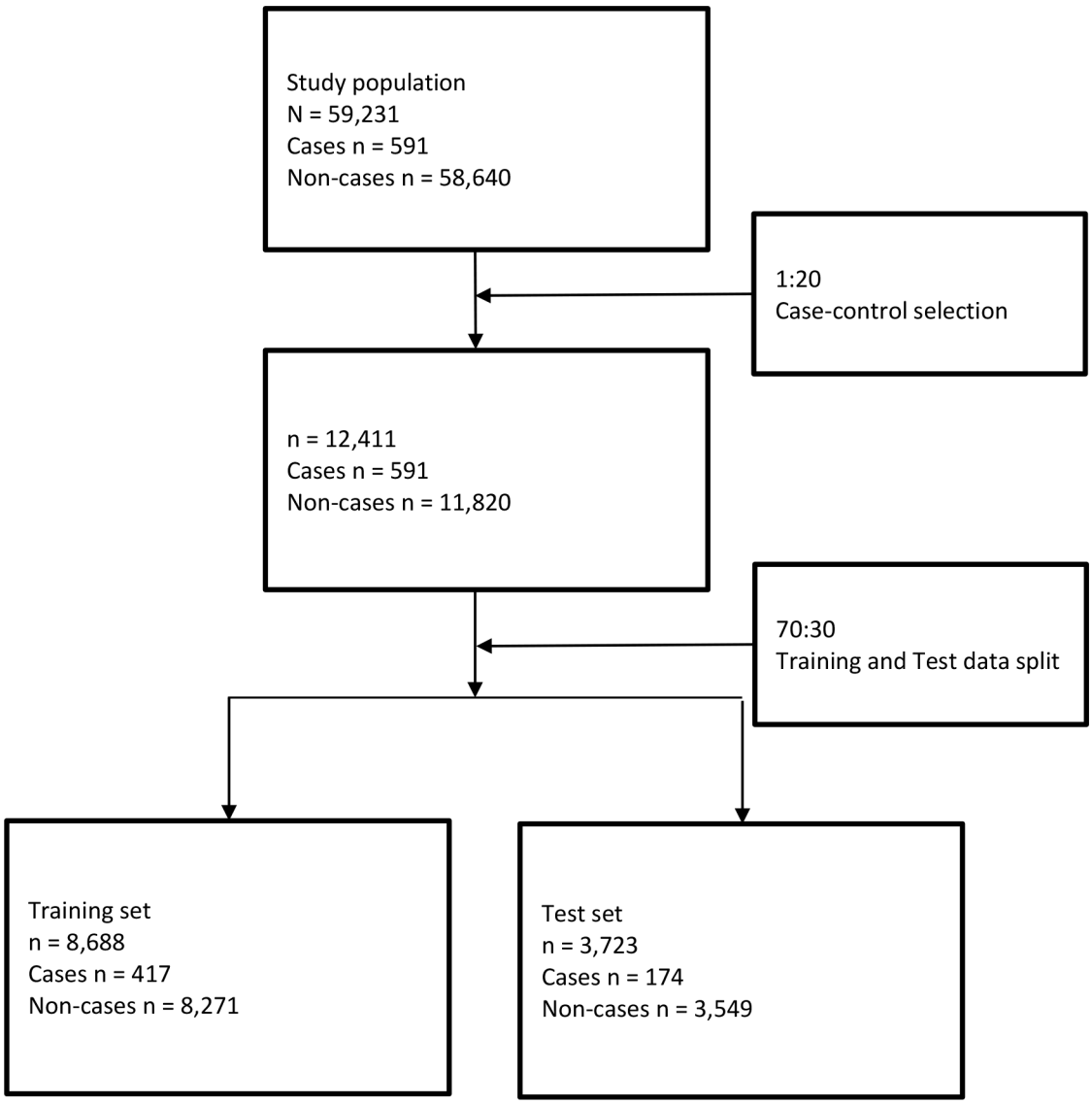
Supplementary Figure 1: Flow diagram of the MCS participants



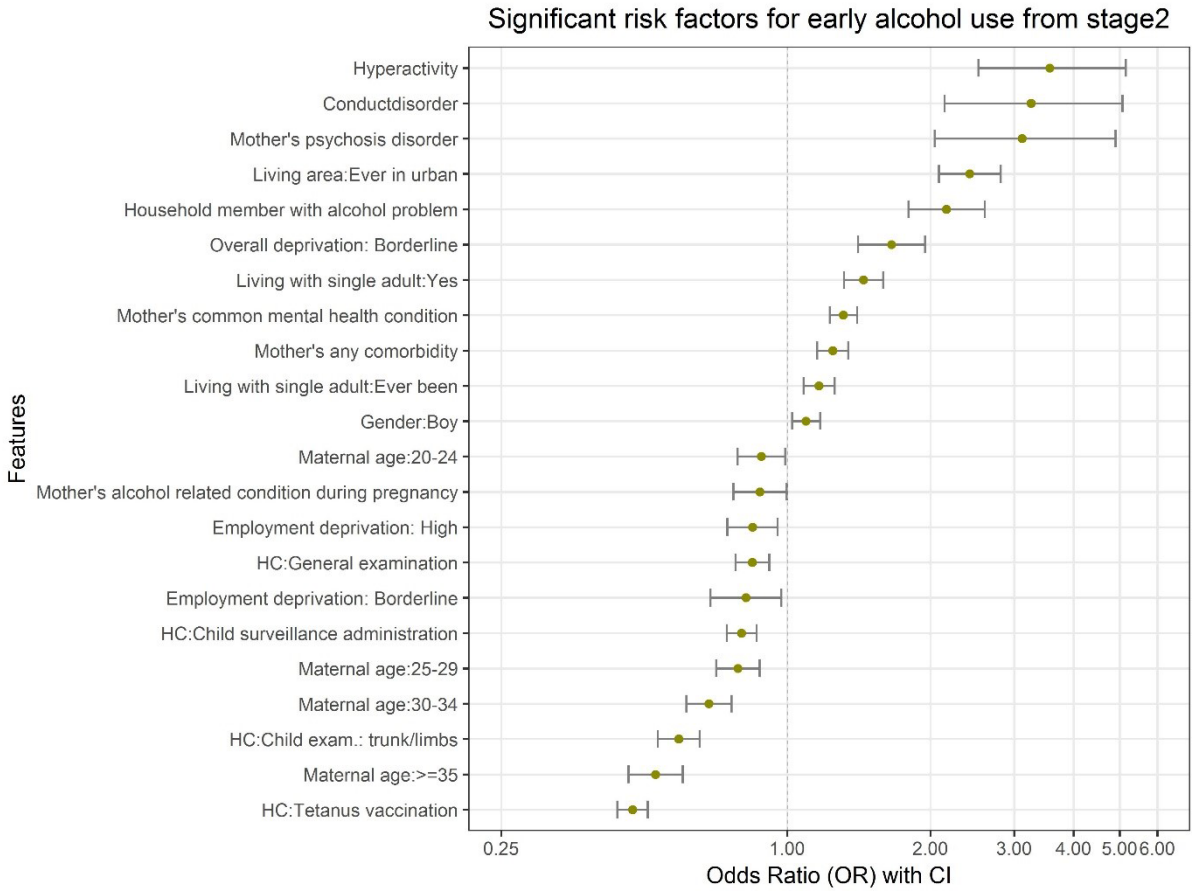
Supplementary Figure 2: Flow diagram of the whole population participants



Supplementary Figure 3: Flow diagram for the final study population



Supplementary Figure 4: Significant risk factors associated with higher and lower risk of early alcohol-related health outcomes from whole population analysis (stage 2)



HC: Health code from EHRs

My input

My contribution to this paper began with developing the methodology to address the research question. To achieve this, I linked all the necessary datasets, including the MCS and other administrative datasets mentioned in the paper, within the SAIL Databank using SQL. In stage two, I constructed the analogous variables in R to address the more complex and granular nature of the time-varying exposure variables derived from routine data. I conducted analyses for both stages in R, which included developing feature selection and predictive regression models. Based on the findings, I authored and published this journal article as the first author.

Impact

- This paper was published in *The International Journal of Population Data Science* in 2022, indicating its contribution to a respected field and enhancing its visibility among researchers.
- This study has been cited by other published works, as noted in Google Scholar, which demonstrates the relevance and impact of this work on the subsequent research.

Conclusion

This study identifies key household and health-related determinants that contribute to child's risk-taking behaviour. Findings suggest that stable household conditions, parental engagement and early healthcare interactions play a significant role in not developing risk-taking behaviours. Targeted interventions to enhance family support systems and access to healthcare could improve long-term developmental outcomes for children at risk. Building on the understanding of childhood environments and risk-taking behaviours, the next chapter will explore how early behavioural difficulties influence adolescent injury risk. The risk profile provided by the current study includes children's externalising behavioural problems, and very granular levels of child behavioural data at different ages are available in the MCS, which will be utilised in the next study. This study has established the strength of incorporating survey data with routine data. Building on this foundation, the next chapter will employ a similar hybridisation of data, which would otherwise be unavailable when relying solely on routine data

Chapter 7: Behavioural difficulties in early childhood and risk of adolescent injury

Critical summary

Background

Injury is one of the leading causes of child mortality (88) and significantly heightens the risk of disability (89). Economic deprivation is closely associated to a higher risk of injury, exacerbating existing inequalities (89). Children with behavioural difficulties, such as Attention Deficit Hyperactive Disorder (ADHD) or Conduct Problem(CP), are more susceptible to injury (90). However, limited evidence exists regarding the long-term association between these behavioural challenges and injury risk. Additionally, much of the existing research relies on self- or proxy-reports to estimate injury risk (91), which may be influenced by recall bias. This study addresses these gaps by investigating the relationship between early childhood behavioural difficulties and long-term injury risk. It examines early behavioural difficulties, specifically ADHD and CP. The research explores how these challenges may compromise a child's safety and increase their vulnerability to injuries during adolescence, a period characterised by greater independence and risk-taking behaviours. Understanding this connection is crucial for developing targeted interventions.

Utilisation of administrative data

This study included MCS data from participants in Wales and Scotland. Parent-reported mental health records of the children, measured using the SDQ (92), were used to assess the children's behavioural difficulties. A data linkage was conducted between the MCS data (with parental consent) and healthcare datasets from Scotland and Wales to capture injury-related hospital admissions and ED visits during adolescence. To integrate health data from both Scotland and Wales, extensive data harmonisation was necessary to align healthcare records from both countries. This process was essential for obtaining the outcome variable, which comprised aggregated frequencies of injury-related hospital admissions and ED visits during the observation period for the study population. This research not only highlights the importance of integrating diverse healthcare datasets but also underscores the critical need for standardised data practices across regions.

Application of data science method

A negative binomial regression model was implemented to account for the over-dispersed outcome data, where the conditional variance was double that of the conditional mean. This model introduces an additional parameter to handle the extra variability, making it particularly suitable for over-dispersed data (93) and providing more accurate estimates of the association between behavioural difficulties and injury risk.

Given the use of survey data, a weight-adjusted model was developed to address oversampling and attrition among participants who did or did not consent to data linkage. This adjustment reduced the potential for selection bias and increased the generalisability of the findings to the wider population. This rigorous methodological approach involved the implementation of appropriate model selection and the use of weight-adjusted modelling for enhancing the validation of the findings.

Early-life vulnerability profiling

The study found no statistically significant long-term association between high behavioural difficulties (hyperactivity and CP) and injury risk after accounting for confounding factors such as gender, demographic characteristics, family factors and socio-economic conditions. This likely reflects the effectiveness of treatment, particularly for those with clinical/high behavioural difficulty, which may reduce injury risk over time. However, the study also found that children identified with borderline CP at an earlier stage were at a higher risk of injury in later periods. This may indicate inadequate support for children with borderline issues, leading to a deterioration of the condition over time. These findings are essential for identifying early intervention opportunities to mitigate the risks of adolescent injuries in at-risk populations.

Published journal paper

OPEN ACCESS

Behavioural difficulties in early childhood and risk of adolescent injury

Amrita Bandyopadhyay ,[1,2] Karen Tingay,[3] Ashley Akbari,[2,4] Lucy Griffiths,[4,5] Helen Bedford ,[5] Mario Cortina-Borja,[6] Suzanne Walton,[5] Carol Dezateux,[7] Ronan A Lyons,[1,2,4] Sinead Brophy[1,2,4]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/archdischild-2019-317271).

[1]National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Swansea, United Kingdom
[2]Administrative Data Research UK, Swansea University Medical School, Swansea, United Kingdom
[3]Office for National Statistics, Cardiff Road, Newport, Wales, UK
[4]Health Data Research UK, Swansea University Medical School, Swansea, United Kingdom
[5]Life Course Epidemiology and Biostatistics, UCL Great Ormond Street Institute of Child Health, UCL, London, UK
[6]Clinical Epidemiology, Nutrition and Biostatistics, UCL Great Ormond Street Institute of Child Health, London, UK
[7]Centre for Primary Care and Public Health, Queen Mary University of London, London, UK

Correspondence to
Amrita Bandyopadhyay, Swansea University Medical School, Swansea SA2 8PP, UK; [redacted]

Received 21 March 2019
Revised 30 September 2019
Accepted 9 October 2019
Published Online First 30 October 2019

Check for updates

To cite: Bandyopadhyay A, Tingay K, Akbari A, *et al*. *Arch Dis Child* 2020;**105**:282–287.

ABSTRACT

Objective To evaluate long-term associations between early childhood hyperactivity and conduct problems (CP), measured using Strengths and Difficulties Questionnaire (SDQ) and risk of injury in early adolescence.

Design Data linkage between a longitudinal birth cohort and routinely collected electronic health records.

Setting Consenting Millennium Cohort Study (MCS) participants residing in Wales and Scotland.

Patients 3119 children who participated in the age 5 MCS interview.

Main outcome measures Children with parent-reported SDQ scores were linked with hospital admission and Accident & Emergency (A&E) department records for injuries between ages 9 and 14 years. Negative binomial regression models adjusting for number of people in the household, lone parent, residential area, household poverty, maternal age and academic qualification, child sex, physical activity level and country of interview were fitted in the models.

Results 46% of children attended A&E or were admitted to hospital for injury, and 11% had high/abnormal scores for hyperactivity and CP. High/abnormal or borderline hyperactivity were not significantly associated with risk of injury, incidence rate ratio (IRR) with 95% CI of the high/abnormal and borderline were 0.92 (95% CI 0.74 to 1.14) and 1.16 (95% CI 0.88 to 1.52), respectively. Children with borderline CP had higher injury rates compared with those without CP (IRR 1.31, 95% CI 1.09 to 1.57).

Conclusions Children with high/abnormal hyperactivity or CP scores were not at increased risk of injury; however, those with borderline CP had higher injury rates. Further research is needed to understand if those with difficulties receive treatment and support, which may reduce the likelihood of injuries.

What is already known on this topic?

- Childhood injury is a leading cause of avoidable mortality and morbidity and disproportionately affects children from disadvantaged backgrounds.
- Children with behavioural difficulties have an increased immediate injury risk.

What this study adds?

- This longitudinal data linkage study found no association between high levels of behavioural difficulties in early childhood and risk of injury in early adolescence.
- Children with borderline conduct problems are at higher long-term injury risk.
- Further work is needed to delineate the persistence of behavioural difficulties through childhood and their relation to support received and subsequent injury risk.

INTRODUCTION

Injury is the leading cause of mortality and ill-health in adolescence.[1 2] It is more common among children from disadvantaged backgrounds and hence contributes to health inequalities.[3] Every year around two million of the overall injury-related visits to Accident & Emergency (A&E) departments involve children and young people in the UK.[4] Investigating modifiable factors associated with increased risk of injury is important to inform appropriate prevention strategies.

Boys, and children involved in higher level of physical activity, in families with a higher poverty level and of younger mothers are at increased risk of childhood injury.[5–8] Attention-deficit hyperactive disorder (ADHD) is one of the most common neuropsychiatric disorders of childhood, with an incidence of 3% to 7% in school-aged children.[9] It is characterised by a higher level of hyperactivity, impulsivity and inattention; children with ADHD are known to be accident-prone with almost a twofold increased risk of injury than children without ADHD.[10] ADHD is diagnosed according to the core symptoms appearing in the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders).[11] Treatment comprises medication such as stimulants or selective norepinephrine reuptake inhibitors such as atomoxetine along with parent training programmes and cognitive–behavioural therapy. Children with ADHD are often unable to estimate the risks associated with their activities, elevating their risk of injuries. Previous research has suggested that hyperactivity measured by the Strengths and Difficulties Questionnaire (SDQ) is associated with an increased risk of unintentional injury.[12 13] Conduct problems (CP) is another externalising behavioural difficulty, which can be conceptualised as antisocial, defiant, aggressive and criminal behavioural pattern in children, which can elevate their injury risk.[11 12] The overall lifetime

prevalence of CP is estimated at 9.5%,[14] but in school-aged children it is around 3%[15] and it is twice as prevalent in boys than girls. Treatment for CP is focused on parent training, family therapy, school behavioural supports with medication prescribed only if treating coexisting ADHD. Previous research suggests that CP is not associated with injury risk after adjusting for ADHD.[16]

To date, the risk association between behavioural difficulty and injury has only been investigated over relatively short periods of follow-up[13 17–20] and it is unclear whether reported associations persist. Previous research also largely relied on self/proxy reports of injury, rather than objective records of injury such as A&E, general practice (GP) or hospital admission (HA) records,[13 17–20] which might be affected by recall bias.[21] We aimed to explore the relationship between early childhood behaviour (as measured by the SDQ), at the age of 5 years, and the risk of injury in early adolescence to identify children who might be at increased risk of future injury.

METHODS

Sample

Data were analysed from the Millennium Cohort Study (MCS), a UK-wide nationally representative longitudinal birth cohort of 18 819 singleton children born between September 2000 and January 2002.[22] Parents were first interviewed in the home when their child was around 9 months old, with subsequent interviews held at 3, 5, 7 and 11 years of age. At age 7, parents gave written consent to link MCS records to their child's routine health records up to their 14th birthday. There were 13 681 singleton children who participated in the age 7 survey, 1951 and 1598 of whom were living in Wales and Scotland, respectively, at the time of interview. Consent for health record linkage was obtained for 3304 singleton children (1839 from Wales and 1465 from Scotland).

Linked cohort

Consented singleton children who participated in the third MCS survey at age 5 years were linked anonymously to their HAs and the A&E department attendances occurring between the ages of 9 and 14 years. The anonymised MCS birth cohort was linked to routinely collected health datasets stored within the privacy protecting Secure Anonymised Information Linkage (SAIL) Databank.[23 24] The linkage procedure has been described in detail elsewhere.[25] Of 3304 consented singleton children, 3269 were linked with their electronic health records (EHRs), of whom 1838 were from Wales and 1431 from Scotland. The study population comprised 3119 children who participated in the age 5 survey in Wales and Scotland and were linked to their health data (figure 1).

Exposure variables

Child behavioural difficulties were assessed by parent-reported SDQ when children were 5 years old. The SDQ is an internationally validated and widely used screening tool to measure child and adolescent behavioural and emotional difficulties.[26 27] In this study, two SDQ subscales were examined: hyperactivity and inattention (restless/overactive, constantly fidgeting, easily distracted, cannot stop and think before acting, lack of attention span) and CP (often has temper, tantrums, disobedience, fights with/bullies other children, often lies or cheats, steals). Each subscale has scores between 0 and 10 with higher scores indicating greater level of difficulty. In this study, we used Goodman's proposed categorisation to assign children into one of three groups: 'normal' (hyperactivity: 0–5; CP: 0–2;), 'borderline' (hyperactivity: 6; CP: 3) or 'high/abnormal' (hyperactivity: 7–10; CP: 4–10).[28] In this study, analyses were performed for all children who had valid hyperactivity (n=3095) and CP (n=3100) scores.

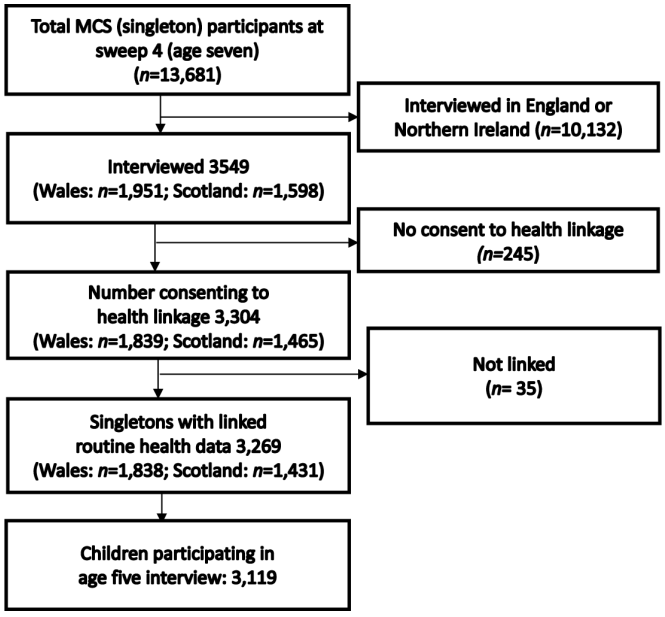


Figure 1 Flow chart of participants. MCS, Millennium Cohort Study.

Outcome variable

The outcome variable was the frequency of injury-related HAs and/or A&E attendances occurring between the age of 9 and 14 years. We identified HA from the Patient Episode Database for Wales (PEDW) and Scottish Morbidity Records dataset for Wales and Scotland, respectively. An injury-related inpatient HA was identified from emergency admissions with an injury diagnosis International Classification of Diseases, version 10 (ICD-10) code appearing in the first diagnostic position, indicating that injury was the primary cause of admission.[29] ICD-10 codes are provided in online supplementary appendix table A1.

We identified A&E attendances from the Emergency Department dataset (EDDS) and the Scottish Accident and Emergency version 2 (A&E2) dataset for Wales and Scotland respectively. Data were collected in EDDS from 2009 and in A&E2 from 2007 onwards. To harmonise the data from both countries, the study focused on A&E attendances recorded from 2009. In addition to the ICD-10 codes, the alphanumeric treatment and diagnosis codes which map to an injury type have been included in the study[30 31] to identify the A&E attendances (refer to online supplementary appendix tables A2 and A3). In this study, simultaneous presence of a patient in both harmonised HA and A&E datasets on the same day, which presumably indicates the transfer from A&E to hospital, has been considered as one record. The data were available as aggregated frequencies of injury-related HAs or A&E attendance per child between ages 9 and 14.

Covariates

The covariates considered here as confounders in the associations between child behavioural difficulties and injury risk include child's sex, maternal age at child's birth, maternal highest academic and vocational qualification (derived by National Vocational Qualifications standard), lone parent carer, household poverty (household income less than 60% of national median using modified Organisation for Economic Co-operation

and Development scale), number of people in the household, residential area (using 2005 Rural/Urban Area Classification) and the child's physical activity level (number of days per week they were involved in sports/exercise). With the exception of sex and maternal age at child's birth (which were collected from age 9 months MCS interview data), these covariates were derived from age 5 MCS interview data.

Statistical analysis

We used negative binomial regression models, as the outcome variable was overdispersed, as indicated by the conditional variances being at least twice the conditional means across all categories of both hyperactivity and CP (see online supplementary appendices 4 and 5). For each SDQ scale, 'normal' was the reference group compared with the 'borderline' and 'high/abnormal' groups in each model. We adjusted these models for the covariates described in the previous section (model 1) and considered models with hyperactivity and CP as main exposure variables as well as a covariate (model 2). We did not contrast these two models in terms of goodness-of-fit criteria, as we were interested principally in the results from model 2, whereas model 1 offered a comparator in terms of the possible changes in the risk measure when the other exposure variable was included as a covariate. Sex-stratified models (models 3 and 4) for each exposure variable were also fitted due to higher levels of hyperactivity and CP in boys and higher levels of injury in boys in general. The models' results were parameterised using incidence rate ratios (IRRs) with 95% CIs. Data preparation including extraction, cleaning and linkage were performed in Structured Query Language on IBM DB2 platform, with all statistical analyses performed in R version 3.3.2.[32] All the models' parameters were estimated adjusting for survey and non-response consent weights to account for oversampling, attrition between consent and non-consent to data linkage in the MCS. Survey and non-response consent weights were obtained via predicted probabilities obtained from logistic regression models taking the stratified cluster sampling design into account and adjusting for low representation of children from Wales, Scotland and Northern Ireland, disadvantaged areas and areas with high proportions of ethnic minority groups. The detailed methodological approach to derive the weight variable has been explained elsewhere.[33]

RESULTS

Table 1 shows the demographic characteristics of the 3119 consented singleton children who participated in the third survey.

Between ages 9 and 14 years, around 46% children had at least one HA or A&E attendance for injury (table 2). There were 2904 records of injury of which 6% were HAs and 94% were A&E attendances.

Associations between SDQ scores and the risk of injury

Hyperactivity at age 5 was not associated with a higher risk of injury in adolescence after adjusting for confounders (table 3). Unadjusted borderline hyperactivity, rather than high/abnormal, were associated with injury (IRR=1.34, 95% CI 1.01 to 1.77). However, this risk attenuated after adjusting for confounding variables, and this effect was similar when CP was included as a covariate.

High/abnormal CP at age 5 were not associated with higher risk of injury in adolescence once adjustment was made for confounding factors (sex, demographic, family factors and socioeconomic variables). However, borderline CP were associated with a higher rate of injury, even when adjusting for confounding factors and hyperactivity (IRR 1.31, 95% CI 1.10 to 1.57). This was especially true for girls with borderline CP (IRR 1.37, 95% CI 1.04 to 1.8), see table 4.

Table 1 Characteristics of all children with linked EHRs in the study population

	n=3119* (weighted %)†
Hyperactivity	
Normal	2563 (82.8)
Borderline	204 (6.5)
High/abnormal	328 (10.6)
Conduct Problems	
Normal	2404 (76.1)
Borderline	391 (13.4)
High/abnormal	305 (10.5)
Sex	
Boy	1604 (51.3)
Girl	1515 (48.7)
No of people in the household	
2	169 (6.4)
3	564 (18.3)
4	1420 (43.9)
five or more	966 (31.4)
Lone parent	
Lone	586 (20.4)
Non-lone	2533 (79.6)
Residential area	
Rural	756 (26.1)
Urban	2360 (73.9)
Household poverty	
OECD 60% median or above	2213 (69.1)
Below OECD 60% median	894 (30.9)
Maternal education	
Degree	635 (19.3)
Diplomas in Higher Education	344 (9.9)
Advanced/Advanced Subsidiary/Subsidiary levels	401 (14.4)
Ordinary level/General Certificate of Secondary Education	1248 (40.0)
Other	77 (2.4)
None	407 (14.0)
Physical activity level	
3 or more days a week	411 (14.6)
2 days a week	581 (18.5)
1 day a week	872 (27.7)
Less often or not at all	1246 (39.1)
Country	
Wales	1750 (36.3)
Scotland	1369 (63.7)
	Median (IQR)
Maternal age at birth (years)	29 (24–33)

*Missing data: hyperactivity (24, unweighted %=0.8); CP (19, unweighted %=0.6); living area (less than 5); poverty indicator (12, unweighted %=0.4); maternal education (7, unweighted %=0.2); physical activity level (9, unweighted %=0.3); maternal age at birth (less than 5).
†Weighting was based on 1 minus the probability of non-consent, multiplied by the age 7 survey weights for data linkage and scaled to the sum of the number of consenting children.
CP, conduct problem; EHR, electronic health record; OECD, Organisation for Economic Co-operation and Development.

Table 2 Number of children with injury records and number of injury-related hospital admissions or A&E attendance records combined by country

Country	No of children with injury (out of n=3119)	Injury admissions/attendances between 9 and 14 years
Wales	866	1782 (HA=104; A&E=1678)
Scotland	562	1122 (HA=70; A&E=1052)
Total	1428 (45.78%)	2904 (HA=174; A&E=2730)

A&E, Accident & Emergency; HA, hospital admission.

DISCUSSION

Our findings suggest that children with high hyperactivity and CP do not have an increased risk of subsequent injury in early adolescence. However, borderline CP, especially in girls, were associated with a greater risk of injury in adolescence. These findings differ from existing studies,[12 13 19] which suggested a significant association between hyperactivity, CP and risk of injury. It is possible that these difficulties do not persist throughout childhood, reflecting either spontaneous resolution or, potentially, the effects of interventions, which include pharmacological or cognitive–behavioural treatments[34] and this may reduce the subsequent injury risk over time. On the other hand, children with borderline disruptive behaviours may not get equivalent family support, parent training/family therapy and school behavioural support. Clinical guidelines for children with behavioural problems have been shown to be inconsistent and difficult to implement due to high caseloads, time pressure and lack of specialised staff.[35] This may result in children with borderline problems not receiving adequate support with potential implications for persistence or worsening of their problems.

Strengths and limitations

A strength of the current study is the longitudinal linkage between routinely collected EHRs with longitudinal survey data[25] which allowed us to examine prospectively recorded injury occurring between 4 and 9 years following assessment of behavioural difficulties. In the current study, we used objective measures of injury and were able to include a longer period of follow-up of the participants, thereby overcoming some of the limitations of previous studies.

However, our study does rely on parent-reported SDQ data. It is possible that parent-reported behaviours may reflect parental perceptions and their ability to cope with child behaviours and so may be subject to bias, for example, parents who are less able to cope or mothers with psychological distress may overestimate their child's behavioural difficulties.[20 36] This may explain why some high/abnormal children were not at risk of injury as their difficulties could have been overestimated by parents. Comparison with teacher assessment would have helped to validate the exposure SDQ measures.

Finally, due to the unavailability of the A&E data prior to 2009, the study included the injury records of participants' between the ages of 9 and 14. Intervention before age 9 due to high rate of injury can reduce the subsequent injury risk, hence data prior to 2009 would have enabled us to investigate the mediating effect of early injury history on the injury risk in adolescence. In this study, GP data were not included; as we did not have GP data for participants from Scotland, we considered that A&E attendances were less likely to include 'worried well' parents. This study considered the first diagnostic code within PEDW to identify the cause of admission and disregarded the secondary or other diagnostic positions to avoid the inclusion of pre-existing comorbidities. This might underestimate some injuries, which were wrongly placed when recorded. In this study, the data were available at an aggregated level per child and the time to injury was not taken into consideration within the current study design. Hence we were not able to distinguish between children with many injuries over a short time frame compared with those who have had them over a longer period. In this study, the missing data for the behavioural difficulties were excluded from the analysis; however, due to the small amount of the missing data, the impact of this exclusion is negligible. Additionally, the study does not capture any injuries that do not result in any healthcare contacts; therefore, the observed association between the behavioural difficulties and injury may have been underestimated.

CONCLUSION

We found no evidence that high/abnormal levels of hyperactivity or CP at school entry are associated with injury risk in later childhood/early adolescence. There is some evidence that borderline CP is associated with injury risk, especially for girls. Children identified as having significant hyperactivity or CP might have received early support or treatment mitigating their risk of long-term injury in adolescence. However, those

Table 3 IRRs for association between hyperactivity and CPs at age 5 and subsequent hospital admissions or A&E attendances for injury between ages 9 and 14 years

	Unadjusted		Adjusted (model 1)*		Adjusted (model 2)†	
	IRR (95% CI)	P value	IRR (95% CI)	P value	IRR (95% CI)	P value
Hyperactivity						
Normal	1		1		1	
Borderline	1.34 (1.01 to 1.77)	0.040	1.21 (0.92 to 1.59)	0.169	1.16 (0.88 to 1.52)	0.299
High	1.10 (0.91 to 1.34)	0.321	0.98 (0.81 to 1.19)	0.837	0.92 (0.74 to 1.14)	0.435
CPs						
Normal	1		1		1	
Borderline	1.38 (1.16 to 1.65)	<0.001	1.31 (1.09 to 1.59)	0.003	1.31 (1.10 to 1.57)	0.003
High	1.26 (1.02 to 1.56)	0.029	1.12 (0.90 to 1.39)	0.307	1.12 (0.89 to 1.42)	0.326

*Model 1: Adjusted for child's sex, number of people in household, lone parent, residential area, household poverty, maternal education, physical activity level, maternal age at child's birth and country of the respondents.
†Model 2: Adjusted for child's sex, number of people in household, lone parent, residential area, household poverty, maternal education, physical activity level, maternal age at child's birth, country of the respondents, CP (in case of hyperactivity) and hyperactivity (in case of CP).
A&E, Accident & Emergency; CP, conduct problem; IRR, incidence rate ratio .

Table 4 IRRs for association between hyperactivity and conduct problems at age 5 and subsequent hospital admissions or A&E attendances for injury between ages 9 and 14 years: sex-stratified models

	Boys				Girls			
	Adjusted (model 3)*		Adjusted (model 4)†		Adjusted (model 3)*		Adjusted (model 4)†	
	IRR (95% CI)	P value	IRR (95% CI)	P value	IRR (95% CI)	P value	IRR (95% CI)	P value
Hyperactivity								
Normal	1		1		1		1	
Borderline	1.05 (0.72 to 1.53)	0.810	1.00 (0.70 to 1.44)	0.982	1.42 (0.91 to 2.21)	0.119	1.33 (0.86 to 2.06)	0.194
High	0.98 (0.77 to 1.25)	0.888	0.93 (0.72 to 1.20)	0.571	0.99 (0.70 to 1.39)	0.937	0.91 (0.62 to 1.33)	0.618
Conduct problems								
Normal	1		1		1		1	
Borderline	1.27 (0.97 to 1.66)	0.088	1.28 (0.98 to 1.67)	0.068	1.38 (1.06 to 1.81)	0.017	1.37 (1.04 to 1.80)	0.023
High	1.07 (0.83 to 1.38)	0.602	1.10 (0.85 to 1.42)	0.465	1.24 (0.81 to 1.91)	0.318	1.20 (0.78 to 1.85)	0.396

*Model 3:Adjusted for number of people in the household, lone parent, residential area, household poverty, maternal education, physical activity level, maternal age at child's birth and country of the respondents.
†Model 4: Adjusted for number of people in the household, lone parent, residential area, household poverty, maternal education, physical activity level, maternal age at child's birth and country of the respondents and also adjusted for conduct problems (in case of hyperactivity) and hyperactivity (in case of conduct problems).
A&E, Accident & Emergency; IRR, incidence rate ratio.

with borderline problems may also be at risk but do not receive necessary support thus maintaining their risk of injuries. Further research is needed to clarify the relation of interventions to behavioural trajectories in early childhood and to investigate whether this modifies future injury risk.

Twitter Karen Tingay @residl_deviance

Acknowledgements The authors are grateful to the Centre for Longitudinal Studies, UCL Institute of Education, NHS Information Standards Division and the UK Data Service. This study makes use of anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank.

Contributors RAL and SB designed the study. AB prepared and analysed the data within the SAIL Databank with support from Mario Cortina-Borja, Karen Tingay and Lucy Griffiths. AB and SB wrote the manuscript and all authors contributed to critically appraising and reviewing the manuscript. All authors approved the final manuscript.

Funding This work was supported by the Wellcome Trust (grant no 087389/B/08/Z), Farr Institute of Health Informatics Research from the Medical Research Council (MR/K006584/1 and MR/K006525/1), Health Data Research UK (grant ref: NIWA1), Administrative Data Research – UK (ES/S007393/1), National Centre for Population Health and Wellbeing Research and Asthma UK Centre for Applied Research (AUK-AC-2012-01).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Ethics approval for the fourth survey of the MCS was received from the Northern and Yorkshire Research Ethics Committee (07/MRE03/32). This study was approved by the SAIL Information Governance Review Panel in Wales and the Public Benefit and Privacy Panel for Health and Social Care in Scotland.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository.

ORCID iDs
Amrita Bandyopadhyay http://orcid.org/0000-0003-2798-4030
Helen Bedford http://orcid.org/0000-0003-0908-1380

REFERENCES

1 Deal LW, Gomby DS, Zippiroli L, *et al*. Unintentional injuries in childhood: analysis and recommendations. *Future Child* 2000;10:4–22.
2 Prevention WHOD of I & V. *The injury Chartbook: a graphical overview of the global burden of injuries*. Geneva: World Health Organization, 2003.
3 Orton E, Kendrick D, West J, *et al*. Persistence of health inequalities in childhood injury in the UK; a population-based cohort study of children under 5. *PLoS One* 2014;9:e111631.
4 National Institute for Health and Clinical Excellence. Unintentional injuries: prevention strategies for under 15S, 2010. Available: https://www.nice.org.uk/Guidance/PH29 [Accessed 4 Jul 2018].
5 Cheng TL, Fields CB, Brenner RA, *et al*. Sports injuries: an important cause of morbidity in urban youth. *Pediatrics* 2000;105:e32.
6 Cubbin C, Smith GS. Socioeconomic inequalities in injury: critical issues in design and analysis. *Annu Rev Public Health* 2002;23:349–75.
7 Orton E, Kendrick D, West J, *et al*. Independent risk factors for injury in pre-school children: three population-based nested case-control studies using routine primary care data. *PLoS One* 2012;7:e35193.
8 Villalba-Cota J, Trujillo-Hernández B, Vásquez C, *et al*. Causes of accidents in children aged 0–14 years and risk factors related to the family environment. *Ann Trop Paediatr* 2004;24:53–7.
9 Cormier E. Attention deficit/hyperactivity disorder: a review and update. *J Pediatr Nurs* 2008;23:345–57.
10 Barkley RA. Attention-deficit hyperactivity disorder. In: *A Handbook for diagnosis and treatment*. 4th edn. Guilford Publications, 2014.
11 American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5TM*. 5th edn. Arlington: American Psychiatric Publishing, Inc, 2013.
12 Lalloo R, Sheiham A. Risk factors for childhood major and minor head and other injuries in a nationally representative sample. *Injury* 2003;34:261–6.
13 Keyes KM, Susser E, Pilowsky DJ, *et al*. The health consequences of child mental health problems and parenting styles: unintentional injuries among European schoolchildren. *Prev Med* 2014;67:182–8.
14 Nock MK, Kazdin AE, Hiripi EVA, *et al*. Prevalence, subtypes, and correlates of DSM-IV conduct disorder in the National comorbidity survey replication. *Psychol Med* 2006;36:699–710.
15 Fairchild G, Hawes DJ, Frick PJ, *et al*. Conduct disorder. *Nat Rev Dis Primers* 2019;5.
16 Schwebel DC, Roth DL, Elliott MN, *et al*. Association of Externalizing behavior disorder symptoms and injury among fifth graders. *Acad Pediatr* 2011;11:427–31.
17 Bijur P*et al*. Behavioral predictors of injury in school-age children. *Arch Pediatr Adolesc Med* 1988;142:1307–12.
18 Rowe R, Maughan B, Goodman R, *et al*. Childhood psychiatric disorder and unintentional injury: findings from a national cohort study. *J Pediatr Psychol* 2004;29:119–30.
19 Lalloo R, Sheiham A, Nazroo JY. Behavioural characteristics and accidents: findings from the health survey for England, 1997. *Accid Anal Prev* 2003;35:661–7.
20 Constant A, Dulioust J, Wazana A, *et al*. Utility of self-reported mental health measures for preventing unintentional injury: results from a cross-sectional study among French schoolchildren. *BMC Pediatr* 2014;14:2.
21 Harel Y, Overpeck MD, Jones DH, *et al*. The effects of recall on estimating annual nonfatal injury rates for children and adolescents. *Am J Public Health* 1994;84:599–605.
22 Connelly R, Platt L. Cohort profile: UK millennium cohort study (mcs). *Int J Epidemiol* 2014;43:1719–25.
23 Ford DV, Jones KH, Verplancke J-P, *et al*. The Sail Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009;9:157.
24 Lyons RA, Jones KH, John G, *et al*. The Sail databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 2009;9:3.

25 Tingay KS, Bandyopadhyay A, Griffiths L, *et al*. Linking consented cohort and routinely collected health data to enhance investigations into childhood obesity, asthma, infections, immunisations, and injuries. *Int J Popul Data Sci* 2019;4.
26 Croft S, Stride C, Maughan B, *et al*. Validity of the strengths and difficulties questionnaire in preschool-aged children. *Pediatrics* 2015;135:e1210–9.
27 Goodman A, Goodman R. Strengths and difficulties questionnaire as a dimensional measure of child mental health. *J Am Acad Child Adolesc Psychiatry* 2009;48:400–3.
28 Goodman R. Scoring the Strengths and Difficulties Questionnaire for age 4–17. *Zugriff Am* 2014;17.
29 World Health Organization. *International statistical classification of diseases and related health problems, 10th revision (ICD-10)*. Geneva: World Health Organ, 1992. http://www.who.int/classifications/icd/en/. (accessed 11 Jun 2018).
30 Lyons RA, Turner S, Lyons J, *et al*. All Wales Injury Surveillance System revised: development of a population-based system to evaluate single-level and multilevel interventions. *Inj Prev* 2016;22:i50–5.
31 NHS Wales Data Dictionary. NHS Wales data dictionary. accident and emergency diagnosis types, 2017. Available: http://www.datadictionary.wales.nhs.uk/#!WordDocuments/accidentandemergencydiagnosistypes.htm [Accessed 20 Nov 2018].
32 R Core Team. R: a language and environment for statistical computing, 2018. Available: https://www.r-project.org/ [Accessed 22 Nov 2018].
33 Sera F, Griffiths L, Dezateux C, *et al*. Technical report on the enhancement of Millennium Cohort Study data with linked electronic health recordsTechnical report on the enhancement of Millennium Cohort Study data with linked electronic health records; derivation of consent weights 2018.
34 Abikoff H, Klein RG. Attention-deficit hyperactivity and conduct disorder: comorbidity and implications for treatment. *J Consult Clin Psychol* 1992;60:881–92.
35 Gatej A-R, Lamers A, van Domburgh L, *et al*. Perspectives on clinical guidelines for severe behavioural problems in children across Europe: a qualitative study with mental health clinicians. *Eur Child Adolesc Psychiatry*;57.
36 Fergusson DM, Lynskey MT, Horwood LJ. The effect of maternal depression on maternal ratings of child behavior. *J Abnorm Child Psychol* 1993;21:245–69.

My input

In this research project, I was involved in building linked datasets between consented MCS data and health and administrative datasets. I undertook extensive and challenging data harmonisation work to incorporate hospital and ED data from both Scotland and Wales, ensuring compatibility for analysis. Since the MCS is a massive source of data, I identified the appropriate exposure and confounding variables from the MCS surveys and linked them anonymously with healthcare records, this was a significant undertaking in this research. Additionally, I developed weight-adjusted regression models to investigate the research question. I am the first author of this research paper and handled correspondence during the publication process.

Impact

- This paper was published in *The Archives of Diseases in Childhood* in 2020, a prestigious journal recognised for its contributions to paediatric research.
- The paper has garnered attention and has been cited by five published works, as reported by Google Scholar. The citations reflect the significance and relevance of the current research work in this field and demonstrates its influence on subsequent research.

Conclusion

This chapter investigates the long-term association between early childhood behavioural difficulties and injury risk in adolescence using a longitudinal data linkage study. While no significant link was found between severe early hyperactivity and long-term injury risk, the findings suggest that children with behavioural challenges are more likely to experience injuries later. This makes them more at risk for injury-related severe outcomes. Additionally, this increased risk can lead to further vulnerabilities, including a higher susceptibility to both physical and mental health challenges. Hence the children with behavioural challenges may still require additional monitoring and tailored interventions to enhance their safety and wellbeing. Misclassification of behavioural difficulties can lead to severe outcomes and increase a child's vulnerability. Given the potential for misclassification, the next chapter examines the risk of ADHD overdiagnosis in Scotland and Wales, exploring the influence of a child's relative age within the school year on ADHD diagnoses and treatment.

Chapter 8: Age within schoolyear and attention-deficit hyperactivity disorder in Scotland and Wales

Critical summary

Background

ADHD is one of the most prevalent neuro-developmental disorders among children and adolescents (94). A recent study reported a significant increase in ADHD diagnoses among children in the U.S., raising considerable public health concerns (95). However, there is variability in the prevalence of ADHD (diagnosed or treated) between countries. While the global prevalence stands at 7.2%, it reaches 15.5% in the U.S. and 23% in Canada (96). Studies have also reported higher prevalence of ADHD among younger children within an academic year (97,98), suggesting a potential risk of over-diagnosis based on relative age.

This study aims to investigate the role of relative age in ADHD diagnoses in the UK, specifically in Wales and Scotland. The UK has a lower prevalence of ADHD compared to many other countries (99), and the school entry cut-off dates differ between Wales (September) and Scotland (March). By comparing these two regions, this research provides valuable insights into the impact of relative age on ADHD diagnosis and its association with confounding factors, including gender, family dynamics and socio-economic conditions.

Utilisation of administrative data

In this study, data from Wales and Scotland were analysed separately. A nationally representative cohort of singleton children with valid education data from Wales between 2009 and 2016 was established, enabling meaningful comparisons with the Scottish cohort. These cohorts were linked to primary care health records in their respective countries to obtain information on ADHD treatment.

Routine administrative datasets were used to capture the exposure, outcome and confounding variables for the analysis. In Wales, the WDS, WLGP, NCCHD and Pre-16 education data were utilised, while in Scotland, the Scottish Morbidity Record maternity database (SMR02) and the ScotXed school pupil census were employed. Data linkage in Wales was performed using the ALF, whereas in Scotland, the Community Health Index (CHI) was used for health records and the Scottish Candidate Number (SCN) for education data linkage.

This study emphasises the strength of administrative data in conducting cross-country comparisons, enhancing the generalisability of the findings. By integrating these comprehensive datasets, the research provides valuable insights into the prevalence and

diagnosis of ADHD at population level, contributing to evidence-based public health strategies.

As this work was a collaborative initiative between Glasgow University and Swansea University, utilising the existing data infrastructure of both countries, only data from Wales and Scotland was incorporated into the study. Data from other two countries of the UK (England and Northern Ireland) was beyond the scope of the study.

Application of data science methods

Both unadjusted and adjusted LR models were applied to examine the relationship between the exposure and outcome variables, as the outcome variable was binary (ADHD treatment: yes/no). When constructing the adjusted models (referred to as Model 2 in the paper), child-related variables, including gender, age within school years and area-level deprivation indicated by the WIMD and the Scottish Index of Multiple Deprivation (SIMD) were included. The final model also adjusted for additional maternal confounding factors to assess their impact on the odds of an ADHD diagnosis.

Although the school entry cut-off dates differ between Scotland and Wales, the analysis was made comparable by deriving four categories of the main exposure variable (age category in school years) based on the month of birth. Chi-square tests were used to compare the demographic characteristics of the study population, both with and without ADHD, ensuring the relevance of including the confounding factors in the final adjusted models. A supplementary analysis was conducted to compare children held back for one year with those in their expected year, using chi-square tests.

Early-life vulnerability profiling

This study revealed very similar prevalence rates of treated ADHD in Wales and Scotland, consistent with findings from other countries such as Denmark and Finland (99). Key findings include, a) the relative age effect is linked to a higher risk of ADHD in Wales; b) a similar trend was observed in Scotland after including held-back children in the analysis; and c) a greater proportion of Scottish children are in the held-back group compared to their Welsh counterparts in an academic year.

These findings suggest that the relative age effect on ADHD should be considered in the clinical context, as younger children within their academic year are more likely to receive ADHD treatment, irrespective of school entry cut-off dates. This highlights the need to consider flexible school start dates in policy decisions. The study significantly contributes to early-life vulnerability profiling, highlighting the necessity for targeted interventions and informed decision-making in ADHD treatment and support.

Published journal paper

BMC Public Health

RESEARCH

Open Access

Age within schoolyear and attention-deficit hyperactivity disorder in Scotland and Wales

Michael Fleming[1*†], Amrita Bandyopadhyay[2,3†], James S. McLay[4], David Clark[5], Albert King[6], Daniel F. Mackay[1], Ronan A. Lyons[2,3,7], Kapil Sayal[8†], Sinead Brophy[2,3,7†] and Jill P. Pell[1†]

Abstract

Background: Previous studies suggest an association between age within schoolyear and attention-deficit hyperactivity disorder (ADHD). Scotland and Wales have different school entry cut-off dates (six months apart) and policies on holding back children. We aim to investigate the association between relative age and treated attention deficit hyperactivity disorder (ADHD) in two countries, accounting for held-back children.

Methods: Routine education and health records of 1,063,256 primary and secondary schoolchildren in Scotland (2009–2013) and Wales (2009–2016) were linked. Logistic regression was used to examine the relationships between age within schoolyear and treated ADHD, adjusting for child, maternity and obstetric confounders.

Results: Amongst children in their expected school year, 8,721 (0.87%) had treated ADHD (Scotland 0.84%; Wales 0.96%). In Wales, ADHD increased with decreasing age (youngest quartile, adjusted OR 1.32, 95% CI 1.19–1.46) but, in Scotland, it did not differ between the youngest and oldest quartiles. Including held-back children in analysis of their expected year, the overall prevalence of treated ADHD was 0.93%, and increased across age quartiles in both countries. More children were held back in Scotland (57,979; 7.66%) than Wales (2,401; 0.78%). Held-back children were more likely to have treated ADHD (Scotland OR 2.18, 95% CI 2.01–2.36; Wales OR 1.70, 95% CI 1.21–2.31) and 81.18% of held-back children would have been in the youngest quartile of their expected year.

Conclusions: Children younger within schoolyear are more likely to be treated for ADHD, suggesting immaturity may influence diagnosis. However, these children are more likely to be held back in countries that permit flexibility, attenuating the relative age effect.

Keywords: Attention-deficit hyperactivity disorder, Relative age, Data linkage, Children, Education, School

What is known on this subject?

- A number of studies have reported that attention-deficit hyperactivity disorder (ADHD) is associated with younger age within schoolyear (relative age) with most investigators suggesting that this may reflect differential case ascertainment, rather than a causal relationship, because younger children are developmentally less mature than their older classmates with whom their behaviour is being compared. Further research has been recommended on the impact of holding back children on case ascertainment and particularly whether the provision of flexibility in school starting dates masks or reduces the relative age effect.

†Michael Fleming and Amrita Bandyopadhyay are joint first authors.

†Kapil Sayal, Sinead Brophy and Jill P. Pell are joint senior authors.

*Correspondence: [redacted]

[1] Institute of Health and Wellbeing, University of Glasgow, 1 Lilybank Gardens, Glasgow G12 8RZ, UK
Full list of author information is available at the end of the article

What this study adds?

- In Scotland and Wales, cut-off ages for school-entry are six months out of phase. They also employ different approaches to holding back children, with the practice being less restrictive in Scotland. Comparison of the two countries, therefore, provides a useful natural experiment for investigating the relationship between age within schoolyear and ADHD, and investigating whether it is independent of potential confounders and modified by policies on holding back children. Clinicians assessing or treating children and young people for ADHD should be aware that irrespective of the date of cut-off for school entry, children who are younger within their school year are more likely to be treated for ADHD. This trend may be masked in countries with flexible start date policies where younger children with attention or behavioural problems are more likely to be held back a year if the teachers and parents agree that this is in the best interests of the child. Holding back children does not appear to reverse the need for ADHD medication. It is possible that holding back children with ADHD might, nonetheless, improve other outcomes.

Background

Attention-deficit hyperactivity disorder (ADHD) is a neuro-developmental disorder, characterised by developmentally inappropriate levels of inattention, hyperactivity and impulsiveness. Diagnosis is based on these symptoms affecting the child's functioning across different settings. Therefore, reports from parents and teachers are considered, in addition to clinical observation and assessment, when making the diagnosis. However, the prevalence of diagnosed or treated ADHD varies greatly between countries, ranging from around 1% in Denmark to over 5% in North America and Iceland [1, 2]. A number of studies have reported that ADHD is associated with younger age within schoolyear (relative age) with most investigators suggesting that this may reflect differential case ascertainment, rather than a causal relationship, because younger children are developmentally less mature than their older classmates with whom their behaviour is being compared [2, 3]. Therefore, younger children may have their immature behaviour misclassified as ADHD (over-ascertainment in younger children) and/or ADHD may not be recognised in some older children who are better able to compensate (more complete ascertainment in younger children). Evidence in support of more complete ascertainment or over-ascertainment among younger children, rather than a genuinely higher incidence in younger children, comes from different sources. Firstly, findings of an association with relative age within schoolyear are more consistent in countries with high ADHD prescription rates, such as the USA, Canada, Iceland, Israel and Germany (pooled risk ratio of 1.27) [2]. In contrast, due to very high heterogeneity, a meta-analysis could not be performed on studies conducted in countries with lower prescription rates [2]. Secondly, in USA states with a fixed 1stSeptember school entry cut-off date, rates of ADHD treatment for young children differed between the youngest and oldest children (i.e. those born in September and August); these differences were not found in states that applied different cut-off dates [4]. However, in common with the vast majority of the literature [3], both this study and a recent UK study [5] were hampered by being unable to identify children held back a year and therefore misclassifying their relative age using month of birth. lt has been suggested that the association with relative age may, therefore, have been underestimated because of exposure misclassification [3, 6]. Further research has been recommended on the impact of holding back children on case ascertainment and particularly whether the provision of flexibility in school starting dates masks or reduces the relative age effect [3]. Thirdly, there are only very modest differences in the likelihood of ADHD treatment receipt between children who are young and old within their schoolyear in Denmark [7], where there are reported to be tight, age-specific criteria for diagnosing ADHD and there is considerable parental discretion in deciding to hold back children considered too immature to start school [8]. In Denmark, 40% of children born October-December, who would normally be in the youngest quartile in their year, are held back to the next year [7].

It is unclear how these international findings might generalise to countries within the UK, such as Scotland and Wales, where healthcare is provided free of charge at the point of delivery via the National Health Service, there is clear guidance for the diagnosis and treatment of ADHD [9, 10], and diagnosis and treatment rates are low [1]. In Scotland and Wales, cut-off ages for school-entry are six months out of phase. They also employ different approaches to holding back children, with the practice being less restrictive in Scotland [11]. Comparison of the two countries, therefore, provides a useful natural experiment for investigating the relationship between age within schoolyear and ADHD, and investigating whether it is independent of potential confounders and modified by policies on holding back children.

Methods

The study was conducted across Scotland (population 5.4million) and Wales (population 3.1million). The two countries are six months out of phase in relation to age at school entry. In Scotland, the cut-off date of birth for entry into school is 1st March and in Wales it is 1st September. Both Scotland and Wales have country-wide coverage of routine health and education data that are linkable at individual-level. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. NHS ethics approval is not required for secondary analyses of anonymised extracts of routine data. Access to the Welsh data was carried out under Secure Anonymised Information Linkage (SAIL) Databank Information Governance Review Panel (IGRP) approved project *Wales Electronic Cohort for Children (WECC): Phase 4* (project number 0916). In Scotland, access was approved by the National Health Service Public Benefit and Privacy Panel (reference 1920–0144) and covered by a data processing agreement between Glasgow University and Public Health Scotland and a data sharing agreement between Glasgow University and the education department of the Scottish Government (ScotXed). All Scottish data were linked by the Electronic Data Research and Innovation Service (eDRIS), part of Public Health Scotland.

Record linkage

In Scotland, the health sector uses a unique identifier, the community health index (CHI), which enables different health databases to be linked to each other, at an individual level, using exact matching. The education sector uses a different unique identifier, the Scottish Candidate Number (SCN), by which different education databases can be linked. We have previously demonstrated that, for singleton children, probabilistic matching of the CHI and SCN, based on date of birth, sex and postcode of residence, is 99% accurate [12]. The linked data were analysed within a secure National Safe Haven. In Wales, data linkage is performed using an anonymised, encrypted NHS number identifier, known as the anonymised linkage field (ALF), which is generated by the trusted third-party NHS Wales Informatics Service (NWIS). The data linkage and analyses were performed within the SAIL Databank platform [13].

Inclusion and exclusion criteria

The study was restricted to singleton children who attended mainstream primary or secondary schools in Scotland between 2009 and 2013 and in Wales between 2009 and 2016, and who had been born in the same country and could, therefore, be linked to their maternity records. Multiple births and special schools were excluded because, in Scotland, it is not possible to be certain that records of same sex children have been correctly linked and school stage is not recorded for special schools. In Wales, children who did not have general practice (GP) records in SAIL could not be included. Since the focus of this study was the effect of age within school year, we excluded children who had been advanced or held back by one or more school year from the primary analyses. This is because they were likely to be either atypically gifted or struggling with academic work independent of their age within year. Therefore, their inclusion would have introduced bias. In the supplementary analyses, we re-included children who had been held back by one year only because, due to less stringent restrictions in Scotland, the parents of children who would otherwise be among the youngest in their year sometimes elect to hold them back a year at entry into primary or secondary school to optimise their exam grades. Since this is not done in response to concerns about the child's academic abilities, their inclusion was less likely to introduce bias.

Databases

In Scotland, the Scottish Morbidity Record maternity database (SMR02) collects data on maternal, obstetric and child factors relating to pregnancy and delivery. The Prescribing Information System (PIS) collects information on all prescriptions dispensed to Scottish residents by community pharmacies or primary care, and includes prescriptions issued in Scotland but dispensed elsewhere in the United Kingdom. The ScotXed school pupil census is conducted annually, in September, by all local authority run primary and secondary schools. In Wales, the study population was derived from the Welsh Demographic Service (WDS) dataset, which is an administrative database of all individuals living in Wales and registered with a GP. Demographic and maternity data were obtained from the WDS and National Community Child Health Database (NCCHD). Medication history was derived from the Welsh Longitudinal General Practice (WLGP) dataset and education records were obtained from the pre-16 years Educational Attainment Dataset.

Exposure, outcome and confounder variables

Age within school year was defined using month of birth and was categorised into calendar (three month) quartiles. Oldest to youngest age within schoolyear equated to births in March–May, June–August, September–November and December-February in Scotland, and September–November, December-February, March–May and June–August in Wales. The outcome of interest was

treated ADHD. In both countries, this was ascertained by receipt of one or more medication licensed solely for the treatment of ADHD defined as: methylphenidate hydrochloride, dexamfetamine sulphate, atomoxetine or lisdexamfetamine dimesylate. The potential confounders included child (sex, area-based socioeconomic deprivation), maternal (smoking during pregnancy, age at delivery, parity) and obstetric (gestation at delivery, sex-gestation-specific birth weight centile, caesarean delivery, 5-min APGAR score) factors. Area-based socioeconomic deprivation was measured using the Scottish Index of Multiple Deprivation (SIMD) and Welsh Index of Multiple Deprivation (WIMD) in Scotland and Wales respectively. Both were categorised into general population quintiles for the respective countries.

Statistical analyses

The characteristics of children with and without ADHD were compared using chi-square tests and chi-square tests for trend for categorical and ordinal data respectively. Binary logistic regression models were used to examine the association between age within schoolyear and ADHD univariately, adjusted for child confounders and, finally, adjusted for all confounders. In the supplementary analyses, the characteristics of children who had been held back one year were compared with children who were in their expected year using chi-square tests and chi-square tests for trend for categorical and ordinal data respectively. The main unadjusted and adjusted binary logistic regression models were then rerun including children who had been held back one year in addition to those in their expected year. Analyses were undertaken using R v3.3.2 and Stata MP version 14.1.

Results

For the primary analyses, the study population comprised 1,002,876 singleton children who were in the expected school year for their age and who attended school in the same country in which they were born. Overall, 8,721 (0.87%) were being treated for ADHD: 5,803 (0.84%) of the 699,325 Scottish schoolchildren and 2,918 (0.96%) of the 303,551 Welsh schoolchildren. The prevalence of ADHD was higher in boys, increased with deprivation, maternal smoking during pregnancy and lower maternal age, birth weight and APGAR score, and had a reverse hockey-stick relationship with gestation at delivery (Table 1).

In Scotland, the prevalence of ADHD increased from the oldest quartile to the second youngest, but then fell in the youngest quartile (Table 2). In Wales, children in the youngest quartile had the highest prevalence of ADHD (Table 2). Differences between Scotland and Wales persisted following adjustment for child, maternal and obstetric confounders (Table 3). In Wales, in the fully adjusted model, the risk of ADHD increased as age decreased over the four quartiles. In Scotland, it increased steadily over the oldest three but the risk of ADHD among children in the lowest age quartile was not significantly higher than in the highest age quartile.

For the supplementary analyses, the study population comprised 1,063,256 children who were either in their expected schoolyear or had been held back a year, of whom 9,897 (0.93%) had treated ADHD. The prevalence of treated ADHD was 0.92% in Scotland and 0.97% in Wales ($p=0.012$) (Supplementary Table 1). If the children who were held back a year had been in their expected schoolyear, the prevalence of ADHD in Scotland across the four quartiles from oldest to youngest would have been: 0.79%, 0.89%, 0.97% and 1.01% (chi trend, $p<0.001$).

More children were held back in Scotland than Wales: 57,979 out of 757,304 (7.66%) versus 2,401 out of 305,991 (0.78%) respectively (Supplementary Table 1). Of the 60,380 held back children, 49,017 (81.18%) would have been in the youngest quartile of their expected year, 8,138 (13.48%) in the third oldest, 2,177 (3.81%) in the second oldest, and only 1,078 (1.74%) in the oldest quartile. Children who were held back a year had a two-fold higher prevalence of ADHD treatment: 1.96% in Scotland, 1.71% in Wales, and 1.95% overall (Supplementary Table 1). Held-back children were more likely to be male, affluent, preterm and low birth weight, and less likely to have been born by Caesarean section or have mothers who smoked during pregnancy (Supplementary Table 2). After adjustment for potential confounders, held-back children remained more likely to have treated ADHD: OR 2.18, 95% CI 2.01–2.36 in Scotland and OR 1.70, 95% CI 1.21–2.31 in Wales respectively (Supplementary Table 3).

Discussion

The prevalence of treated ADHD in Scotland and Wales was comparable to countries such as Denmark and Finland, and lower than the USA [1]. When our analyses included only children who were in their expected school year, younger relative age was associated with higher risk of ADHD in Wales, but not in Scotland. Scottish children were ten times more likely than Welsh children to be held back a year. However, the lower prevalence of ADHD in the lowest age quartile in Scotland was explained by preferential holding back of children who were closer to the cut-off age (who would, otherwise, have been amongst the youngest in the year) and preferential holding back of children with treated ADHD. When the analyses included held-back children, and was based on their expected schoolyear, there was a clear trend, in both

Table 1 Characteristics of schoolchildren in their expected school year, by presence or absence of ADHD

	Scotland					Wales					Overall				
	ADHD		No ADHD		*p* value	ADHD		No ADHD		*p* value	ADHD		No ADHD		
	N = 5,803		*N* = 693,522			*N* = 2,918		*N* = 300,633			*N* = 8,721		*N* = 994,155		
	n	%	n	%		n	%	n	%		n	%	n	%	
Gender					< 0.001[a]					< 0.001[a]					
Male	4,921	84.80	342,819	49.43		2,469	84.61	153,229	50.97		7,390	84.74	496,048	49.90	
Female	882	15.20	350,703	50.57		449	15.39	147,404	49.03		1,331	15.26	498,107	50.10	
Welsh/Scottish Index of Multiple Deprivation					< 0.001[b]					< 0.001[b]					
1 (Most deprived)	1,863	32.18	157,444	22.73		987	33.82	73,646	24.50		2,850	32.73	231,090	23.27	
2	1,457	25.17	139,997	20.22		667	22.86	62,204	20.69		2,124	24.39	202,201	20.36	
3	1,015	17.53	133,947	19.34		537	18.40	59,051	19.64		1,552	17.82	192,998	19.43	
4	856	14.79	134,187	19.38		398	13.64	48,846	16.25		1,254	14.40	183,033	18.43	
5 (Least deprived)	598	10.33	126,962	18.33		329	11.27	56,886	18.92		927	10.65	183,848	18.51	
Missing	14		985			0		0			14		985		
Maternal age (years)					< 0.001[b]					< 0.001[b]					
19	1,005	17.32	56,408	8.13		478	16.38	27,893	9.28		1,483	17.00	84,301	8.48	
20–24	1,680	28.95	133,302	19.22		910	31.19	67,145	22.34		2,590	29.70	200,447	20.16	
25–29	1,562	26.92	204,476	29.48		777	26.63	87,434	29.09		2,339	26.82	291,910	29.37	
30–34	1,071	18.46	196,610	28.35		498	17.07	76,870	25.58		1,569	17.99	273,480	27.51	
[3]35	485	8.36	102,714	14.81		255	8.74	41,202	13.71		740	8.49	143,916	14.48	
Missing	0		12			0		89			0		101		
Maternal smoking					< 0.001[a]					< 0.001[a]					
No	2,673	51.86	446,432	72.65		452	58.25	65,806	79.13		3,125	52.70	512,238	73.42	
Yes	2,481	48.14	168,055	27.35		324	41.75	17,352	20.87		2,805	47.30	185,407	26.58	
Missing	649		79,035			2,142		217,475			2,791		296,510		
Gestational age (weeks)					< 0.001[b]					< 0.001[b]					
< 28	21	0.36	721	0.10		17	0.58	643	0.21		38	0.44	1,364	0.14	
28–34	181	3.12	13,478	1.94		99	3.39	6,823	2.27		280	3.21	20,301	2.04	
35–36	315	5.43	23,315	3.36		137	4.69	10,693	3.56		452	5.18	34,008	3.42	
37–41	5,088	87.72	630,151	90.93		2,488	85.26	264,048	87.83		7,576	86.90	894,199	89.99	
[3]42	195	3.36	25,369	3.66		177	6.07	18,426	6.13		372	4.27	43,795	4.41	
Missing	3		488			0		0			3		488		
Birth weight (g)					< 0.001[b]					< 0.001[b]					
£1,000	25	0.43	938	0.14		11	0.38	568	0.19		36	0.41	1,506	0.15	
1,001–1,500	3	0.05	192	0.03		20	0.69	1,498	0.50		23	0.26	1,690	0.17	
1,501–2,500	466	8.03	35,769	5.16		227	7.81	14,722	4.91		693	7.96	50,491	5.08	
2,501–4,000	4,693	80.87	569,582	82.15		2,345	80.72	248,585	82.89		7,038	80.82	818,167	82.37	
4,001–4,500	519	8.94	73,333	10.58		253	8.71	29,356	9.79		772	8.87	102,689	10.34	
> 4,500	97	1.67	13,496	1.95		49	1.69	5,186	1.73		146	1.68	18,682	1.88	
Missing	0		212			13		718			13		930		
5 min Apgar score					< 0.001[b]					< 0.001[b]					
0–3	42	0.74	3,116	0.45		128	7.06	9,641	4.41		170	2.26	12,757	1.41	
4–6	77	1.35	6,305	0.92		24	1.32	1,882	0.86		101	1.34	8,187	0.90	
7–10	5,595	97.92	677,504	98.63		1,662	91.62	206,996	94.73		7,257	96.40	884,500	97.69	
Missing	89		6,597			1,104		82,114			1,193		88,711		
Parity					0.893[a]					< 0.001[a]					
0	2,630	45.73	315,009	45.64		1,134	42.74	114,650	43.87		3,764	44.79	429,659	45.16	
[3]1	3,121	54.27	375,160	54.36		1,519	57.26	146,687	56.13		4,640	55.21	521,847	54.84	
Missing	52		3,353			265		39,296			317		42,649		

Table 1 (continued)

	Scotland					Wales					Overall				
	ADHD		No ADHD		*p* value	ADHD		No ADHD		*p* value	ADHD		No ADHD		
	N = 5,803		*N* = 693,522			*N* = 2,918		*N* = 300,633			*N* = 8,721		*N* = 994,155		
	n	%	n	%		n	%	n	%		n	%	n	%	
Mode of delivery					0.207[a]					0.025[a]					
Caesarean section	1,132	19.51	139,915	20.17		512	17.55	57,686	19.19		5,183	59.43	611,291	61.49	
Other	4,671	80.49	553,605	79.83		2,406	82.45	242,947	80.81		3,538	40.57	382,862	38.51	
Missing	0		2			0		0			0		2		

[a] chi square test for association

[b] chi square test for trend

P values produced separately as cohorts could not be combined

Table 2 Breakdown of ADHD by age within school year

	Scotland*			Wales **		
	Month of birth	N	ADHD N (%)	Month of birth	N	ADHD N (%)
1 (oldest)	Mar-May	186,002	1,450 (0.78)	Sept-Nov	76,944	684 (0.89)
2	June-Aug	191,822	1,637 (0.85)	Dec-Feb	74,430	730 (0.98)
3	Sept-Nov	184,751	1,609 (0.87)	Mar-May	75,655	737 (0.97)
4 (youngest)	Dec-Feb	136,750	1,107 (0.81)	June-Aug	76,552	767 (1.00)

[*] Chi squared test for trend $p = 0.164$

[**] Chi squared test for trend $p = 0.112$

Table 3 Logistic regression models of the association between age within school year and ADHD

Age category within school year	Model 1			Model 2			Model 3		
	OR	95% CI	*p* value	OR	95% CI	*p* value	OR	95% CI	*p* value
Scotland									
1 (oldest)	1.00			1.00			1.00		
2	1.10	1.02–1.18	0.012	1.11	1.03–1.19	0.004	1.12	1.04–1.21	0.002
3	1.12	1.04–1.20	0.002	1.12	1.04–1.20	0.003	1.13	1.05–1.22	0.001
4 (youngest)	1.04	0.96–1.12	0.343	1.06	0.98–1.15	0.151	1.06	0.98–1.15	0.154
Wales									
1 (oldest)	1.00			1.00			1.00		
2	1.10	0.99—1.23	0.063	1.16	1.04—1.28	0.007	1.15	1.03—1.28	0.01
3	1.10	0.99—1.22	0.083	1.21	1.09—1.35	< 0.001	1.20	1.08—1.34	< 0.001
4 (youngest)	1.13	1.02—1.25	0.022	1.32	1.19—1.47	< 0.001	1.32	1.19—1.46	< 0.001

OR Odds ratio, *CI* Confidence interval

Model 1: univariate

Model 2: adjusted for child (sex, age, deprivation quintile) confounders

Model 3; adjusted for above plus, maternal (smoking, age) and obstetric (parity, gestation at delivery, sex- gestation-specific birthweight centile, caesarean section, 5-min Apgar score) confounders

countries, between relative age within schoolyear and treated ADHD.

Children who were held back a year differed from their peers in a number of ways and are likely to be a heterogeneous group. They were more likely to have a range of risk factors for ADHD, including male sex, maternal smoking during pregnancy, preterm birth and low birth weight. However, they were also more likely to

be from affluent families, suggesting that the preferential take-up of school deferral by some families might be driven by parental worry about perceived relative immaturity. Whilst we were unable to examine the reasons for why children were held back, we assume that it reflects a belief that younger children with behavioural or attention problems may fare badly competing against older, more mature peers and/or might benefit from additional schooling.

Our study had a number of strengths. Firstly, it was large-scale (sample size exceeding one million) and non-selective, in that it covered the whole population of both countries. Secondly, linkage with educational databases enabled us to distinguish which children had been held back. This addressed a major limitation of most previous studies that could not and, therefore, systematically misclassified the relative age of held-back children based on their month of delivery [3, 5, 6, 14, 15]. Thirdly, linkage of education to maternity records enabled us to adjust for maternal and obstetric confounders as well as sociodemographic confounders. Fourthly, inclusion of and comparison between two countries permitted a natural experiment in which we could examine whether the relationship between age within schoolyear and ADHD was influenced by school-entry date cut-off and different approaches to permitting children to be held back a year. ADHD is 60–90% heritable [16]. Expression of candidate genes related to the dopamine system, is modified by exposure to sunlight [17], and their association with ADHD interacts with season of birth [18, 19]. However, our finding that younger children were at higher risk of ADHD in two countries, six months out of phase, demonstrated that the association with relative age is not simply due to confounding by month or season of birth.

A previous UK study of one million children attending school in England/Wales, Scotland or Northern Ireland reported an increased overall risk of ADHD among those in the bottom three-month age group (adjusted HR 1.36, 95% CI 1.28–1.45) [5]. However, the overall finding was dominated by children from England/Wales who accounted for more than 90% of the cohort and had the same school entry date; the authors reported that they were underpowered to comment on the Scottish subgroup. Furthermore, the study was conducted using only primary care records. The authors did not have access to education records and, therefore, could not differentiate children in their expected schoolyear from children who had been held back.

Limitations of our study include the use of medication to identify children with ADHD. Although this may have resulted in failure to ascertain some diagnosed cases, there is no reason to believe that it would introduce a systematic error in relation to month of birth. Our study used data extracted from routine administrative databases, but these undergo regular quality assurance checks. Had the children who were held back a year been in their expected school year, the prevalence of ADHD in Scotland would have followed the same pattern as Wales, increasing across the four quartiles from oldest to youngest. Also, the prevalence of treated ADHD would have been comparable in the two countries. In both countries, delaying school entry by a year was associated with a greater, not lower, likelihood of treatment for ADHD. Therefore, holding back younger children for a year did not appear to reverse their need for ADHD medication. However, we were unable to ascertain at what stage in their education the children were held back to a lower year and, therefore, whether commencement of ADHD treatment preceded or followed their deferral.

In terms of clinical and policy implications, clinicians assessing or treating children and young people for ADHD should be aware that irrespective of the date of cut-off for school entry, children who are younger within their school year are more likely to be treated for ADHD. This trend may be masked in countries with flexible start date policies where younger children with attention or behavioural problems are more likely to be held back a year if the teachers and parents agree that this is in the best interests of the child. Holding back children does not appear to reverse the need for ADHD medication. It is possible that holding back children with ADHD might, nonetheless, improve other outcomes. Further studies are required to determine whether holding back children with ADHD produces other benefits such as improvements in their behavioural and educational outcomes and wellbeing.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12889-022-13453-w.

Additional file 1: Table S1. Breakdown of ADHD by age within school year.

Additional file 2: Table S2. Comparison of the characteristics of children held-back one year with those in their expected school year.

Additional file 3: Table S3. Logistic regression models of the association between age within school year and ADHD.

Acknowledgements

The authors would like to acknowledge the support of the electronic Data Research and Innovation Services (eDRIS) within Public Health Scotland for their involvement in obtaining approvals, provisioning and linking Scottish data, and supporting use of the secure analytical platform within the National Safe Haven. The Welsh study used data provided by patients and collected by the NHS Wales as part of their care and support. We analysed anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank and would like to acknowledge all the data providers who enable SAIL to make anonymised data available for research compliance with ethical standards.

Authors' contributions
KS had the original concept. All authors agreed the study design. DC and AK provided Scottish data and undertook record linkage. The Welsh data were provided by SAIL databank and AB undertook the record linkage. MF and AB undertook the statistical analyses. All authors interpreted the results. JPP, MF and AB drafted the manuscript and all other authors contributed revisions. All authors reviewed and approved the final version of the manuscript.

Funding
The Scottish part of the study was sponsored by Health Data Research UK (grant reference number MR/S003800/1). The sponsor and funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review or approval of the manuscript, or decision to submit the manuscript for publication.
This Welsh research was funded as part of the ADR Wales programme of work. The ADR Wales programme of work is aligned to the priority themes as identified in the Welsh Government's national strategy: Prosperity for All. ADR Wales brings together data science experts at Swansea University Medical School, staff from the Wales Institute of Social and Economic Research, Data and Methods (WISERD) at Cardiff University and specialist teams within the Welsh Government to develop new evidence which supports Prosperity for All by using the SAIL Databank at Swansea University, to link and analyse anonymised data. ADR Wales is part of the Economic and Social Research Council (part of UK Research and Innovation) funded ADR UK (grant ES/S007393/1). This Welsh research was also supported by the National Centre for Population Health and Well-Being Research (NCPHWR).

Availability of data and materials
The authors applied for permission to access, link and analyse the Scottish data and undertook mandatory training in data protection, IT security and information governance. All Scottish data were accessed and analysed within a secure national safe haven environment. Therefore, the datasets generated and analysed during the study are not publicly available. The anonymised routinely collected Welsh data accessed in the study were granted via the SAIL databank. The datasets generated and analysed during the current study are available in the SAIL databank repository, https://saildatabank.com/. The availability of the data is subject to request.

Declarations

Ethics approval and consent to participate
All methods were carried out in accordance with relevant guidelines and regulations. The need for ethical approvals to access Scottish and Welsh data were waived by the NHS West of Scotland Research Ethics Service and the SAIL Databank Information Governance Review Panel (IGRP) respectively. The former confirmed that formal NHS ethics approval was not required for the analyses of the Scottish data since the study involved linkage of routinely collected data with an acceptably negligible risk of identification. The Scottish data were approved by the National Health Service Public Benefit and Privacy Panel and covered by a data processing agreement between Glasgow University and Public Health Scotland and a data sharing agreement between Glasgow University and the education department of the Scottish Government (ScotXed). The Welsh data were approved by SAIL Databank Information Governance Review Panel (IGRP) and the data were analysed under the approved project Wales Electronic Cohort for Children (WECC): Phase 4 (project number 0916).
The need for informed consents for Scottish and Welsh data were also waived by the NHS West of Scotland Research Ethics Service and the SAIL Databank Information Governance Review Panel (IGRP) respectively. It was deemed that obtaining informed consent from patients did not apply to our study which involved retrospectively linking and analysing already collected and centrally held routine administrative data comprising anonymised electronic patient records.

Consent for publication
Not applicable.

Competing interests
All authors declare that they have no competing interests.

Author details
[1]Institute of Health and Wellbeing, University of Glasgow, 1 Lilybank Gardens, Glasgow G12 8RZ, UK. [2]Administrative Data Research Wales, Swansea University Medical School, Swansea SA2 8PP, UK. [3]National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Swansea SA2 8PP, UK. [4]Department of Child Health, University of Aberdeen, Aberdeen AB25 2ZG, UK. [5]Public Health Scotland, Edinburgh EH12 9EB, UK. [6]ScotXed, Scottish Government, Edinburgh EH6 6QQ, UK. [7]Health Data Research UK, Swansea University Medical School, Swansea SA2 8PP, UK. [8]Division of Psychiatry & Applied Psychology, University of Nottingham, Nottingham NG7 2UH, UK.

Received: 5 May 2021 Accepted: 3 May 2022
Published online: 30 May 2022

References

1. Sayal K, Prasad V, Daley D, Ford T, Coghill D. ADHD in Children and Young People: Prevalence, Care Pathways, and Service Provision. Lancet Psychiatry. 2018;5(2):175–86.
2. Holland J, Sayal K. Relative age and ADHD symptoms, diagnosis and medication: a systematic review. Eur Child Adolesc Psychiatry. 2019;28(11):1417–29. https://doi.org/10.1007/s00787-018-1229-6.
3. Whitely M, Raven M, Timimi S, et al. Attention deficit hyperactivity disorder late birthdate effect common in both high and low prescribing international jurisdictions: a systematic review. J Child Psychol Psychiatry. 2019;60(4):380–91. https://doi.org/10.1111/jcpp.12991.
4. Layton TJ, Barnett ML, Hicks TR, Jena AB. Attention deficit-hyperactivity disorder and month of school enrollment. N Engl J Med. 2018;379(22):2122–30. https://doi.org/10.1056/NEJMoa1806828.
5. Root A, Brown JP, Forbes HJ, Bhaskaran K, Hayes J, Smeeth L, Douglas IJ. Association of relative age in the school year with diagnosis of intellectual disability, attention-deficit/hyperactivity disorder, and depression. JAMA Ped. 2019;173(11):1068–75.
6. Sayal K, Chudal R, Hinkka-Yli-Salomäki S, Joelsson P, Sourander A. Relative age within the school year and diagnosis of attention-deficit hyperactivity disorder: a nationwide population-based study. Lancet Psychiatry. 2017;4(11):868–75. https://doi.org/10.1016/S2215-0366(17)30394-2.
7. Pottegard A, Hallas J, Zoega H. Children's relative age and use of medication for ADHD: a Danish nationwide study. J Child Psychol. 2014;55(11):1244–50.
8. Dalsgaard S, Humlum MK, Nielsen HS, Simonsen M. Common Danish standards in prescribing medication for children and adolescents with ADHD. Eur Child Adolesc Psych. 2014;23(9):841–4.
9. NICE Guideline [NG87]. Attention deficit hyperactivity disorder: Diagnosis and management. London: National Institute for Health and Care Excellence; 2018.
10. SIGN Guideline [112]. Management of attention deficit and hyperkinetic disorders in children and young people. Edinburgh: Scottish Intercollegiate Guidelines Network; 2009.
11 Bradshaw P, Hall S, Hill T, Mabelis J, Philo D. Growing up in Scotland: early experiences of primary school. Edinburgh: Scottish Government; 2012.
12. Wood R, Clark D, King A, Mackay D, Pell J. Novel cross-sectoral linkage of routine health and education data at an all-Scotland level: a feasibility study. Lancet. 2013;382:S10.
13. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke JP, Ford DV, Brown G, Leake K. The SAIL databank: linking multiple health and social care datasets. BMC Med Inform Decis Mak. 2009;9:3.
14. Hoshen M, Benis A, Keyes K, Zoëga H. Stimulant use for ADHD and relative age in class among children in Israel. Pharmacoepidemiol Drug Saf. 2016;25(6):652–60.
15. Krabbe EE, Thoutenhoofd ED, Conradi M, Pijl S, Batstra L. Birth month as predictor of ADHD medication use in Dutch school classes. Eur J Spec Needs Educ. 2014;29(4):571–8.
16. Thapar A, Holmes J, Poulton K, Harrington R. Genetic basis of attention deficit and hyperactivity. Br J Psychiatry. 1999;174:105–11.
17. Naber D, Wirz-Justice A, Kafka MS. Circadian rhythm in rat brain opiate receptor. Neurosci Lett. 1981;21(1):45–50.
18. Seeger G, Schloss P, Schmidt MH, Ruter-Jungfleisch A, Henn FA. Gene-environment interaction in hyperkinetic conduct disorder (HD + CD)

as indicated by season of birth variations in dopamine receptor (DRD4) gene polymorphism. Neurosci Lett. 2004;366(3):282–6.
19. Brookes KJ, Neale B, Xu X, Thapar A, Gill M, Langley K, et al. Differential dopamine receptor D4 allele association with ADHD dependent of proband season of birth. Am J Med Genet B Neuropsychiatr Genet. 2008;147B(1):94–9.

Publisher's Note

My input

As this study incorporates data from two separate countries, the analyses were conducted independently. The analysis for the Welsh data was performed in SAIL, where I undertook all the necessary steps for data integration, linkage, cleaning and analysis. After completing the analyses on both platforms, we combined the findings. I communicated rigorously with my counterpart in Scotland to harmonise the data and ensure the validity and comparability of the final analysis. I collaborated on writing the paper with my Scottish counterpart, taking the lead on the manuscript's development and incorporating the methods and results from Wales. I also contributed to finalising the introduction and discussion sections of the paper as a joint first author.

Impact

- The paper was published in *BMC Public Health* in 2022, marking a significant contribution to the field.
- It has been cited by eleven other published works, highlighting its substantial impact and relevance in health and social care research.

Conclusion

This chapter of the thesis provides crucial insights into the risk of ADHD overdiagnosis in Wales and Scotland, highlighting its association with relative age within an academic year. The findings indicate that younger children in their academic cohort are more likely to receive an ADHD diagnosis and treatment, raising concerns about the influence of relative age on diagnostic patterns. These results have significant implications for educational policies and clinical decision-making, emphasising the need for greater awareness of relative age effects in ADHD assessments. Considering the next step of this work, federated analysis can open up the possibility for comparative research between countries, like this, to examine how such an approach can effectively explore natural policy and practice differences across the UK and beyond. Together with the previous chapters, this thesis has consistently investigated the risk factors contributing to early-life vulnerabilities through the systematic use of routine administrative health and social care data using statistical and machine learning approaches.

Chapter 9: Conclusion

Summary

The collective findings of this thesis illustrate the multifaceted nature of early-life vulnerabilities, including health disparities, educational challenges, behavioural risks and exposure to DA, as identified through routine administrative data. Across nine published journal articles, my research highlights the critical role of data-informed approaches in identifying early-life vulnerabilities. I have established a methodological framework that uses data linkage and advanced statistical and machine learning techniques on routine administrative data. This framework enhances the identification of early-life vulnerabilities, showcasing the effectiveness of data-driven models. My thesis highlights the necessity of data hybridisation, effectively integrating extensive routine data with detailed survey data through data linkage and harmonisation. This integration enables in-depth longitudinal follow-up of population-level data and facilitates the development of multi-stage models. The methodological approaches have demonstrated how analogous variables from survey data have been created by exploring the complementary strengths of diverse datasets. My thesis showcases significant methodological advancements in handling routine administrative data for health and social care research. The methodological enhancement has culminated in the development of WECC Phase 4, a novel platform which I have developed for life course research and it facilitates future collaborative efforts in this field.

Along with the methodological contribution, my thesis has provided a risk profile associated with early-life vulnerabilities. The risk factors identified by my work include family and area-level deprivation, maternal mental and physical health issues, maternal smoking and clinically significant alcohol or substance abuse by family members. Findings also indicate that children experiencing neglect or challenging family circumstances have fewer routine primary care contacts and more frequent emergency healthcare interactions. The thesis demonstrates that early identification of vulnerable children is possible through integrated linked routine data, opening opportunities for targeted early intervention plans to improve life trajectories. This work also provides the insight into propensity to come out of poverty. For example, living in an area with better community safety and access to services translates into children more likely to build a resilience profile for themselves.

Some of the risk factors associated with early-life vulnerability, such as deprivation, poor mental health, are well known. In this context, one of the contributions of my study is to identify the persistent presence of some of these known modifiable risk factors for vulnerable children in Wales, where social services could help to address the significant underlying disparity or gaps. My findings highlight the need for evidence-based policies to address the root causes of vulnerability, promoting proactive public health

interventions. These are aligned with the priorities of Welsh Government and Public Health Wales (PHW), particularly through my close collaboration with the 'Vulnerability Profiling Programme' of PHW Rhondda Cynon Taf, which aims to improve the life chances of vulnerable children and their families.

Limitations

While this research provides valuable insights, there are areas for improvement in both data utilisation and methodological approaches to investigating early-life vulnerabilities. Primarily, this work relies on the secondary use of administrative data, meaning individuals not captured in the healthcare system are excluded from analysis, potentially omitting important subgroups. Additionally, variability in data quality and completeness may introduce biases. However, substantial efforts, including data cleaning, imputation of missing data and harmonisation, have been undertaken to mitigate these issues. The predictive accuracy of the models can be further improved by expanding the scope of explanatory variables, refining interrelationships and implementing more complex models that offer better fit, though often at the expense of interpretability. Data privacy concerns and governance regulations can restrict accessibility or introduce delays in the research execution. Addressing these challenges requires thoughtful project design that accounts for data availability, governance constraints and methodological rigor. This thesis does not include any patient and public involvement (PPI) engagement. The lack of PPI engagement is a limitation of the study; however, my future research endeavours will engage with PPI groups from the respective field.

Future work

The following areas will be emphasised for developing a future roadmap of this research to enhance its effectiveness and relevance.

- Implementing multilevel modelling will help capture complex hierarchical relationships between exposure variables. This approach enables data analysis at multiple levels: individual, family, community as well as policy-level, providing a more nuanced understanding of how different factors interact and influence health outcomes.
- Expanding machine learning methodologies by incorporating more sophisticated algorithms, such as random forests and neural networks, will enhance predictive capabilities. These advanced techniques can uncover patterns/interactions in large datasets that traditional methods may overlook.
- Establishing strong communication channels with stakeholders, including policymakers and PPI groups, is crucial. Engaging with individuals who have experienced vulnerable situations ensures that the research remains grounded in and is shaped by real-world experiences. I have already started working with existing PPI groups, including stakeholders such as Welsh Women’s Aid, Stories,

Calan, the Early Years Group for WG and the HAPPEN primary school network in Wales. This collaborative approach not only enhances the relevance of the work but also fosters trust and support for data-driven public health initiatives.

These findings highlight exciting opportunities to refine and expand the scope of research questions and methodological approaches, advancing the optimal use of routine administrative data. The future holds great promise for uncovering valuable insights hidden within these vast datasets. Ensuring that these findings are effectively communicated to policymakers will be essential in shaping interventions that improve the life chances of children facing early-life vulnerabilities.

References

1. Mackenzie C, Rogers W, Dodds S, editors. Vulnerability: New Essays in Ethics and Feminist Philosophy [Internet]. Oxford University Press; 2013 [cited 2025 Nov 4]. Available from: https://doi.org/10.1093/acprof:oso/9780199316649.001.0001

2. Alaszewski A. Vulnerablity and risk across the life course. Health Risk Soc [Internet]. 2013 Aug 1 [cited 2025 Nov 4]; Available from: https://www.tandfonline.com/doi/abs/10.1080/13698575.2013.822852

3. Mullin A. Children, Vulnerability, and Emotional Harm. In: Mackenzie C, Rogers W, Dodds S, editors. Vulnerability: New Essays in Ethics and Feminist Philosophy [Internet]. Oxford University Press; 2013 [cited 2025 Nov 4]. p. 0. Available from: https://doi.org/10.1093/acprof:oso/9780199316649.003.0012

4. Jack P. Shonkoff DAP. From Neurons to Neighborhoods: The Science of Early Childhood Development [Internet]. Shonkoff JP, editor. Washington (DC): National Academies Press (US); 2000 [cited 2024 Dec 18]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK225557/

5. Duncan GJ, Dowsett CJ, Claessens A, Magnuson K, Huston AC, Klebanov P, et al. School readiness and later achievement. Dev Psychol. 2007 Nov;43(6):1428–46.

6. James Heckman. The Heckman Equation. 2016 [cited 2024 Dec 18]. Invest in Early Childhood Development: Reduce Deficits, Strengthen the Economy. Available from: https://heckmanequation.org/resource/invest-in-early-childhood-development-reduce-deficits-strengthen-the-economy/

7. McLoyd VC. Socioeconomic disadvantage and child development. Am Psychol. 1998;53(2):185–204.

8. Batko K, Ślęzak A. The use of Big Data Analytics in healthcare. J Big Data [Internet]. 2022 Jan 6 [cited 2025 Feb 1];9(1):3. Available from: https://doi.org/10.1186/s40537-021-00553-4

9. Sarker IH. Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. SN Comput Sci [Internet]. 2021 Jul 12 [cited 2025 Feb 1];2(5):377. Available from: https://doi.org/10.1007/s42979-021-00765-8

10. Choroszewicz M. (In)visible everyday work of fostering a data-driven healthcare and social service organisation. New Technol Work Employ [Internet]. 2024 [cited 2025 Feb 1];39(1):1–18. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/ntwe.12270

11. Kerasidou A, Kerasidou C (Xaroula). Data-driven research and healthcare: public trust, data governance and the NHS. BMC Med Ethics [Internet]. 2023 Jul 14 [cited 2025 Feb 1];24(1):51. Available from: https://doi.org/10.1186/s12910-023-00922-z

12. Balmer S, Black NM, Brown C, Cookson R, Crossley S, Eddy L, et al. An evidence-based plan for addressing poverty with and through education settings. 2024 Mar 15 [cited 2025 Feb 1]; Available from: https://durham-repository.worktribe.com/output/2408020

13. Garavito GAA, Moniz T, Mansilla C, Iqbal S, Dobrogowska R, Bennin F, et al. Activities used by evidence networks to promote evidence-informed decision-making in the health sector– a rapid evidence review. BMC Health Serv Res [Internet]. 2024 Feb 29 [cited 2025 Feb 1];24(1):261. Available from: https://doi.org/10.1186/s12913-024-10744-3

14. Powell L, Spencer S, Clegg J, Wood M. A country that works for all children and young people: An evidence-based approach to supporting children in the preschool years. 2024;

15. Khare SK, March S, Barua PD, Gadre VM, Acharya UR. Application of data fusion for automated detection of children with developmental and mental disorders: A systematic review of the last decade. Inf Fusion [Internet]. 2023 Nov 1 [cited 2025 Feb 1];99:101898. Available from: https://www.sciencedirect.com/science/article/pii/S1566253523002142

16. SAIL databank. Health Data Research Innovation Gateway [Internet]. 2024 [cited 2024 Dec 31]. Available from: https://healthdatagateway.org/en/search?search=&datasetpublisher=SAIL&datasetSort=latest&tab=Datasets&query=EDDS

17. Ford DV, Jones KH, Verplancke JP, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Serv Res [Internet]. 2009 Sep 4 [cited 2017 Nov 29];9:157. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2744675/

18. Jones K, Ford D, Thompson S, Lyons R. A Profile of the SAIL Databank on the UK Secure Research Platform. Int J Popul Data Sci [Internet]. [cited 2022 Jun 9];4(2):1134. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8142954/

19. Jones KH, Ford DV, Jones C, Dsilva R, Thompson S, Brooks CJ, et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: a privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform. 2014 Aug;50(100):196–204.

20. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke JP, Ford DV, et al. The SAIL databank: linking multiple health and social care datasets. BMC Med Inform Decis Mak [Internet]. 2009 Jan 16 [cited 2018 Aug 9];9(1):3. Available from: https://doi.org/10.1186/1472-6947-9-3

21. Hutchings E, Loomes M, Butow P, Boyle FM. A systematic literature review of researchers' and healthcare professionals' attitudes towards the secondary use and sharing of health administrative and clinical trial data. Syst Rev [Internet]. 2020 Oct

12 [cited 2024 Dec 19];9(1):240. Available from: https://doi.org/10.1186/s13643-020-01485-5

22. Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C. Life course epidemiology. J Epidemiol Community Health. 2003;57(10):778.

23. Cuervo H, Cook J. Understanding Young Lives Through Longitudinal Research Design. In: Wyn J, Cahill H, Woodman D, Cuervo H, Leccardi C, Chesters J, editors. Youth and the New Adulthood: Generations of Change [Internet]. Singapore: Springer; 2020 [cited 2024 Dec 30]. p. 13–30. Available from: https://doi.org/10.1007/978-981-15-3365-5_2

24. Hughes RA, Tilling K, Lawlor DA. Combining Longitudinal Data From Different Cohorts to Examine the Life-Course Trajectory. Am J Epidemiol [Internet]. 2021 Dec 1 [cited 2024 Dec 30];190(12):2680–9. Available from: https://doi.org/10.1093/aje/kwab190

25. Hurren E, Stewart A, Dennison S. New Methods to Address Old Challenges: The Use of Administrative Data for Longitudinal Replication Studies of Child Maltreatment. Int J Environ Res Public Health [Internet]. 2017 Sep [cited 2024 Dec 30];14(9):1066. Available from: https://www.mdpi.com/1660-4601/14/9/1066

26. Soneson E, Das S, Burn AM, van Melle M, Anderson JK, Fazel M, et al. Leveraging Administrative Data to Better Understand and Address Child Maltreatment: A Scoping Review of Data Linkage Studies. Child Maltreat [Internet]. 2023 Feb 1 [cited 2025 Jan 9];28(1):176–95. Available from: https://doi.org/10.1177/10775595221079308

27. Saunders NR, Janus M, Porter J, Lu H, Gaskin A, Kalappa G, et al. Use of administrative record linkage to measure medical and social risk factors for early developmental vulnerability in Ontario, Canada. Int J Popul Data Sci [Internet]. 2021 Feb 11 [cited 2025 Jan 9];6(1). Available from: https://ijpds.org/article/view/1407

28. Findlay L, Beasley E, Park J, Kohen D, Algan Y, Vitaro F, et al. Longitudinal child data: What can be gained by linking administrative data and cohort data? Int J Popul Data Sci [Internet]. 2018 Nov 14 [cited 2025 Jan 9];3(1). Available from: https://ijpds.org/article/view/451

29. Cheng C, Messerschmidt L, Bravo I, Waldbauer M, Bhavikatti R, Schenk C, et al. A General Primer for Data Harmonization. Sci Data [Internet]. 2024 Jan 31 [cited 2025 Jan 9];11(1):152. Available from: https://www.nature.com/articles/s41597-024-02956-3

30. Adhikari K, Patten SB, Patel AB, Premji S, Tough S, Letourneau N, et al. Data Harmonization and Data Pooling from Cohort Studies: A Practical Approach for Data Management. Int J Popul Data Sci [Internet]. 2021 Nov 30 [cited 2025 Jan 9];6(1). Available from: https://ijpds.org/article/view/1680

31. Gurugubelli VS, Fang H, Shikany JM, Balkus SV, Rumbut J, Ngo H, et al. A review of harmonization methods for studying dietary patterns. Smart Health Amst Neth. 2022 Mar;23:100263.

32. Fuchs S, Thaler T. Vulnerability and Resilience to Natural Hazards. Cambridge University Press; 2018. 369 p.

33. Garrido EF, Weiler LM, Taussig HN. Adverse Childhood Experiences and Health-Risk Behaviors in Vulnerable Early Adolescents. J Early Adolesc [Internet]. 2018 May 1 [cited 2025 Jan 10];38(5):661–80. Available from: https://doi.org/10.1177/0272431616687671

34. Green MJ, Tzoumakis S, Laurens KR, Dean K, Kariuki M, Harris F, et al. Latent profiles of early developmental vulnerabilities in a New South Wales child population at age 5 years. Aust N Z J Psychiatry [Internet]. 2018 Jun 1 [cited 2025 Jan 10];52(6):530–41. Available from: https://doi.org/10.1177/0004867417740208

35. Curtin M, Madden J, Staines A, Perry IJ. Determinants of vulnerability in early childhood development in Ireland: a cross-sectional study. BMJ Open [Internet]. 2013 Jan 1 [cited 2025 Jan 10];3(5):e002387. Available from: https://bmjopen.bmj.com/content/3/5/e002387

36. Talarico F, Liu YS, Metes D, Wang M, Wearmouth D, Kiyang L, et al. Risk factors for developmental vulnerability: Insight from population-level surveillance using the Early Development Instrument. Digit Health [Internet]. 2023 Nov 3 [cited 2025 Jan 10];9:20552076231210705. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10624014/

37. Rohat G. Projecting Drivers of Human Vulnerability under the Shared Socioeconomic Pathways. Int J Environ Res Public Health [Internet]. 2018 Mar [cited 2025 Jan 10];15(3):554. Available from: https://www.mdpi.com/1660-4601/15/3/554

38. Walsh D, McCartney G, Smith M, Armour G. Relationship between childhood socioeconomic position and adverse childhood experiences (ACEs): a systematic review. J Epidemiol Community Health [Internet]. 2019 Dec 1 [cited 2025 Jan 10];73(12):1087–93. Available from: https://jech.bmj.com/content/73/12/1087

39. Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation [Internet]. 2016 Feb 9 [cited 2025 Jan 10];133(6):601–9. Available from: https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.115.017719

40. Kleinbaum DG, Klein M. Survival Analysis: A Self-Learning Text, Third Edition. Springer; 2011. 708 p.

41. Bayliss LE, Culliford D, Monk AP, Glyn-Jones S, Prieto-Alhambra D, Judge A, et al. The effect of patient age at intervention on risk of implant revision after total replacement of the hip or knee: a population-based cohort study. The Lancet

[Internet]. 2017 Apr 8 [cited 2025 Jan 10];389(10077):1424–30. Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(17)30059-4/fulltext

42. Schober P, Vetter TR. Linear Regression in Medical Research. Anesth Analg [Internet]. 2020 Jan [cited 2025 Jan 10];132(1):108–9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7717471/

43. Zabor EC, Reddy CA, Tendulkar RD, Patil S. Logistic Regression in Clinical Studies. Int J Radiat Oncol Biol Phys [Internet]. 2022 Feb 1 [cited 2025 Jan 10];112(2):271–7. Available from: https://www.redjournal.org/article/S0360-3016(21)02646-8/fulltext

44. Tso WWY, Wong RS, Tung KTS, Rao N, Fu KW, Yam JCS, et al. Vulnerability and resilience in children during the COVID-19 pandemic. Eur Child Adolesc Psychiatry [Internet]. 2022 Jan 1 [cited 2025 Jan 10];31(1):161–76. Available from: https://doi.org/10.1007/s00787-020-01680-8

45. Cao M, Xu D, Xie F, Liu E, Liu S. The influence factors analysis of households' poverty vulnerability in southwest ethnic areas of China based on the hierarchical linear model: A case study of Liangshan Yi autonomous prefecture. Appl Geogr [Internet]. 2016 Jan 1 [cited 2025 Jan 10];66:144–52. Available from: https://www.sciencedirect.com/science/article/pii/S0143622815300217

46. du Toit M, van der Linde J, Swanepoel DW. Early Childhood Development Risks and Protective Factors in Vulnerable Preschool Children from Low-Income Communities in South Africa. J Community Health [Internet]. 2021 Apr 1 [cited 2025 Jan 10];46(2):304–12. Available from: https://doi.org/10.1007/s10900-020-00883-z

47. Santana CLA, Manfrinato CV, Souza PRP, Marino A, Condé VF, Stedefeldt E, et al. Psychological distress, low-income, and socio-economic vulnerability in the COVID-19 pandemic. Public Health [Internet]. 2021 Oct 1 [cited 2025 Jan 10];199:42–5. Available from: https://www.sciencedirect.com/science/article/pii/S0033350621003437

48. Simons A, Govender R, Saunders CJ, Singh-Adriaanse R, Van Niekerk A. Childhood vulnerability to drowning in the Western Cape, South Africa: Risk differences across age and sex. Child Care Health Dev. 2020;46(5):607–16.

49. Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. J Clin Epidemiol [Internet]. 2016 Mar 1 [cited 2025 Jan 10];71:76–85. Available from: https://www.sciencedirect.com/science/article/pii/S0895435615004667

50. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. Fam Med Community Health [Internet]. 2020 Feb 16 [cited 2025 Jan 10];8(1):e000262. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7032893/

51. Percha B. Modern Clinical Text Mining: A Guide and Review. Annu Rev Biomed Data Sci. 2021 Jul 20;4:165–87.

52. Sedlakova J, Daniore P, Horn Wintsch A, Wolf M, Stanikic M, Haag C, et al. Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. PLOS Digit Health. 2023 Oct;2(10):e0000347.

53. Boateng EY, Abaye DA. A Review of the Logistic Regression Model with Emphasis on Medical Research. J Data Anal Inf Process [Internet]. 2019 Oct 12 [cited 2024 Dec 9];07(04):190. Available from: http://www.scirp.org/journal/Paperabs.aspx?PaperID=95655

54. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons; 2013.

55. Loh WY. Fifty Years of Classification and Regression Trees. Int Stat Rev [Internet]. 2014 [cited 2024 Dec 9];82(3):329–48. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12016

56. Zhang H, Singer BH. Recursive partitioning in the health sciences. Springer Science & Business Media; 2013.

57. Burger K. How does early childhood care and education affect cognitive development? An international review of the effects of early interventions for children from different social backgrounds. Early Child Res Q [Internet]. 2010 Apr 1 [cited 2021 Mar 19];25(2):140–65. Available from: https://www.sciencedirect.com/science/article/pii/S0885200609000921

58. Camacho C, Straatmann VS, Day JC, Taylor-Robinson D. Development of a predictive risk model for school readiness at age 3 years using the UK Millennium Cohort Study. BMJ Open [Internet]. 2019 Jun 1 [cited 2021 Mar 18];9(6):e024851. Available from: https://bmjopen.bmj.com/content/9/6/e024851

59. Husa RA, Parrish JW, Johnson HS. Pre-Birth Household Challenges Predict Future Child's School Readiness and Academic Achievement. Children [Internet]. 2022 Mar 15 [cited 2025 Jan 16];9(3):414. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8947585/

60. Kokkalia G, Drigas AS, Economou A, Roussos P. School Readiness From Kindergarten to Primary School. 2019 [cited 2025 Jan 16]; Available from: https://core.ac.uk/outputs/482964889/?source=2

61. Dillman DA. Internet, phone, mail, and mixed-mode surveys : the tailored design method [Internet]. Hoboken : Wiley; 2014 [cited 2025 Jan 16]. 538 p. Available from: http://archive.org/details/internetphonemai0000dill

62. Chen SW, Keglovits M, Devine M, Stark S. Sociodemographic Differences in Respondent Preferences for Survey Formats: Sampling Bias and Potential Threats to

External Validity. Arch Rehabil Res Clin Transl [Internet]. 2022 Mar 1 [cited 2025 Jan 16];4(1):100175. Available from: https://www.sciencedirect.com/science/article/pii/S2590109521000914

63. Moreno-Serra R, Anaya-Montes M, León-Giraldo S, Bernal O. Addressing recall bias in (post-)conflict data collection and analysis: lessons from a large-scale health survey in Colombia. Confl Health [Internet]. 2022 Apr 8 [cited 2025 Jan 16];16(1):14. Available from: https://doi.org/10.1186/s13031-022-00446-0

64. Maier C, Thatcher JB, Grover V, Dwivedi YK. Cross-sectional research: A critical perspective, use cases, and recommendations for IS research. Int J Inf Manag [Internet]. 2023 Jun 1 [cited 2025 Jan 16];70:102625. Available from: https://www.sciencedirect.com/science/article/pii/S0268401223000063

65. Louis D, Oberoi S, Ricci MF, Pylypjuk C, Alvaro R, Seshia M, et al. School Readiness Among Children Born Preterm in Manitoba, Canada. JAMA Pediatr. 2022 Oct 1;176(10):1010–9.

66. Shah PE, Kaciroti N, Richards B, Lumeng JC. Gestational Age and Kindergarten School Readiness in a National Sample of Preterm Infants. J Pediatr [Internet]. 2016 Nov 1 [cited 2021 Mar 24];178:61–7. Available from: https://www.sciencedirect.com/science/article/pii/S0022347616304917

67. Sanz A. The Impact of Poverty on Educational Achievement: Understanding Socio-Economic Barriers and Opportunities for Change. Electron Theses Diss [Internet]. 2024 Jan 1; Available from: https://spark.bethel.edu/etd/1128

68. Ait Belkacem N, Gorgui J, Tchuente V, Aubin D, Lippé S, Bérard A. Maternal Mental Health in Pregnancy and Its Impact on Children's Cognitive Development at 18 Months, during the COVID-19 Pandemic (CONCEPTION Study). J Clin Med [Internet]. 2024 Feb 13 [cited 2025 Jan 19];13(4):1055. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10889100/

69. Mensah FK, Kiernan KE. Maternal general health and children's cognitive development and behaviour in the early years: findings from the Millennium Cohort Study. Child Care Health Dev [Internet]. 2011 [cited 2025 Jan 19];37(1):44–54. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2214.2010.01150.x

70. Phua DY, Kee MZL, Meaney MJ. Positive Maternal Mental Health, Parenting, and Child Development. Biol Psychiatry [Internet]. 2020 Feb 15 [cited 2025 Jan 19];87(4):328–37. Available from: https://www.sciencedirect.com/science/article/pii/S0006322319317780

71. Townsend P, Phillimore P, Beattie A. Health and Deprivation: Inequality and the North. London: Routledge; 2023. 236 p.

72. Visser K, Bolt G, Finkenauer C, Jonker M, Weinberg D, Stevens GWJM. Neighbourhood deprivation effects on young people's mental health and well-being:

A systematic review of the literature. Soc Sci Med [Internet]. 2021 Feb 1 [cited 2025 Jan 22];270:113542. Available from: https://www.sciencedirect.com/science/article/pii/S0277953620307619

73. Ilie S, Sutherland A, Vignoles A. Revisiting free school meal eligibility as a proxy for pupil socio-economic deprivation. Br Educ Res J [Internet]. 2017 [cited 2025 Jan 22];43(2):253–74. Available from: https://www.jstor.org/stable/44954826

74. Welsh Government. GOV.WALES. 2011 [cited 2020 Aug 5]. Welsh Index of Multiple Deprivation (full Index update with ranks): 2011. Available from: https://gov.wales/welsh-index-multiple-deprivation-full-index-update-ranks-2011

75. Petru G. A Review on the Economic Costs of Domestic Violence in the Republic of Moldova. In 2024 [cited 2025 Jan 29]. Available from: https://ibn.idsi.md/vizualizare_articol/202303

76. Walker-Descartes I, Mineo M, Condado LV, Agrawal N. Domestic Violence and Its Effects on Women, Children, and Families. Pediatr Clin [Internet]. 2021 Apr 1 [cited 2025 Jan 29];68(2):455–64. Available from: https://www.pediatric.theclinics.com/article/S0031-3955(20)30183-8/fulltext

77. Doroudchi A, Zarenezhad M, Hosseininezhad H, Malekpour A, Ehsaei Z, Kaboodkhani R, et al. Psychological complications of the children exposed to domestic violence: a systematic review. Egypt J Forensic Sci [Internet]. 2023 May 26 [cited 2025 Jan 29];13(1):26. Available from: https://doi.org/10.1186/s41935-023-00343-4

78. Nesca M, Au W, Turnbull L, Brownell M, Brownridge DA, Urquia ML. Intentional injury and violent death after intimate partner violence. A retrospective matched-cohort study. Prev Med [Internet]. 2021 Aug 1 [cited 2025 Jan 29];149:106616. Available from: https://www.sciencedirect.com/science/article/pii/S0091743521002000

79. Kothari CL, Rhodes KV, Wiley JA, Fink J, Overholt S, Dichter ME, et al. Protection orders protect against assault and injury: A longitudinal study of police-involved women victims of intimate partner violence. J Interpers Violence. 2012;27(14):2845–68.

80. Richards L. Domestic Abuse, Stalking and Harassment and Honour Based Violence (DASH, 2009) Risk Identification and Assessment and Management Model. 2009; Available from: https://reducingtherisk.org.uk/wp-content/uploads/2022/08/DASH-2009.pdf

81. Kim MJ, Mason WA, Herrenkohl TI, Catalano RF, Toumbourou JW, Hemphill SA. Influence of Early Onset of Alcohol Use on the Development of Adolescent Alcohol Problems: A Longitudinal Binational Study. Prev Sci Off J Soc Prev Res [Internet]. 2017 Jan [cited 2025 Jan 23];18(1):1–11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5235962/

82. Kristjansson AL, Santilli AM, Mills R, Layman HM, Smith ML, Mann MJ, et al. Risk and Resilience Pathways, Community Adversity, Decision-making, and Alcohol Use Among Appalachian Adolescents: Protocol for the Longitudinal Young Mountaineer Health Study Cohort. JMIR Res Protoc. 2022 Aug 5;11(8):e40451.

83. Lipperman-Kreda S, Grube JW. Associations of Early Age of First Intoxication with Past Year Drinking Contexts and Problems. Subst Use Misuse. 2019;54(7):1146–53.

84. Donovan JE, Molina BSG. Childhood Risk Factors for Early-Onset Drinking. J Stud Alcohol Drugs [Internet]. 2011 Sep [cited 2025 Feb 5];72(5):741–51. Available from: https://www.jsad.com/doi/abs/10.15288/jsad.2011.72.741

85. Hartmann SA, Hayes T, Sutherland MT, Trucco EM. Risk factors for early use of e-cigarettes and alcohol: Dimensions and profiles of temperament. Dev Psychopathol. 2023 May;35(2):481–93.

86. Sartor CE, Lynskey MT, Heath AC, Jacob T, True W. The role of childhood risk factors in initiation of alcohol use and progression to alcohol dependence. Addiction [Internet]. 2007 [cited 2025 Feb 5];102(2):216–25. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1360-0443.2006.01661.x

87. Connelly R, Platt L. Cohort Profile: UK Millennium Cohort Study (MCS). Int J Epidemiol [Internet]. 2014 Dec 1 [cited 2018 Aug 9];43(6):1719–25. Available from: https://academic.oup.com/ije/article/43/6/1719/703283

88. Gallagher L, Breslin G, Leavey G, Curran E, Rosato M. Determinants of unintentional injuries in preschool age children in high-income countries: A systematic review. Child Care Health Dev. 2024 Jan;50(1):e13161.

89. Jernbro C, Bonander C, Beckman L. The association between disability and unintentional injuries among adolescents in a general education setting: Evidence from a Swedish population-based school survey. Disabil Health J [Internet]. 2020 Jan 1 [cited 2025 Jan 26];13(1):100841. Available from: https://www.sciencedirect.com/science/article/pii/S1936657419301517

90. Ruiz-Goikoetxea M, Cortese S, Aznarez-Sanado M, Magallon S, Luis EO, Zallo NA, et al. Risk of unintentional injuries in children and adolescents with ADHD and the impact of ADHD medications: protocol for a systematic review and meta-analysis. BMJ Open [Internet]. 2017 Sep 1 [cited 2018 Nov 20];7(9):e018027. Available from: https://bmjopen.bmj.com/content/7/9/e018027

91. Keyes KM, Susser E, Pilowsky DJ, Hamilton A, Bitfoi A, Goelitz D, et al. The health consequences of child mental health problems and parenting styles: unintentional injuries among European schoolchildren. Prev Med. 2014 Oct;67:182–8.

92. Goodman A, Goodman R. Strengths and Difficulties Questionnaire as a Dimensional Measure of Child Mental Health. J Am Acad Child Adolesc Psychiatry [Internet]. 2009 Apr 1 [cited 2017 Nov 30];48(4):400–3. Available from: http://www.sciencedirect.com/science/article/pii/S0890856709600472

93. Green JA. Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression. Health Psychol Behav Med [Internet]. 2021 Jan 1 [cited 2025 Jan 27];9(1):436–55. Available from: https://doi.org/10.1080/21642850.2021.1920416

94. Peterson BS, Trampush J, Brown M, Maglione M, Bolshakova M, Rozelle M, et al. Tools for the Diagnosis of ADHD in Children and Adolescents: A Systematic Review. Pediatrics [Internet]. 2024 Mar 25 [cited 2025 Jan 28];153(4):e2024065854. Available from: https://doi.org/10.1542/peds.2024-065854

95. Danielson ML, Claussen AH, Bitsko RH, Katz SM, Newsome K, Blumberg SJ, et al. ADHD Prevalence Among U.S. Children and Adolescents in 2022: Diagnosis, Severity, Co-Occurring Disorders, and Treatment. J Clin Child Adolesc Psychol Off J Soc Clin Child Adolesc Psychol Am Psychol Assoc Div 53. 2024;53(3):343–60.

96. Gascon A, Gamache D, St-Laurent D, Stipanicic A. Do we over-diagnose ADHD in North America? A critical review and clinical recommendations. J Clin Psychol [Internet]. 2022 [cited 2025 Jan 28];78(12):2363–80. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jclp.23348

97. Holland J, Sayal K. Relative age and ADHD symptoms, diagnosis and medication: a systematic review. Eur Child Adolesc Psychiatry [Internet]. 2019 Nov 1 [cited 2025 Jan 28];28(11):1417–29. Available from: https://doi.org/10.1007/s00787-018-1229-6

98. Sayal K, Chudal R, Hinkka-Yli-Salomäki S, Joelsson P, Sourander A. Relative age within the school year and diagnosis of attention-deficit hyperactivity disorder: a nationwide population-based study. Lancet Psychiatry [Internet]. 2017 Nov 1 [cited 2025 Jan 28];4(11):868–75. Available from: https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(17)30394-2/abstract

99. Sayal K, Prasad V, Daley D, Ford T, Coghill D. ADHD in children and young people: prevalence, care pathways, and service provision. Lancet Psychiatry [Internet]. 2018 Feb 1 [cited 2025 Jan 28];5(2):175–86. Available from: https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(17)30167-0/abstract

Appendices

Supplementary paper 01

International Journal of Population Data Science

Journal Website: www.ijpds.org

Record linkage to enhance consented cohort and routinely collected health data from a UK birth cohort

Tingay, K.S.[1*†], Bandyopadhyay, A.[2], Griffiths, L.[2], Akbari, A.[3], Brophy, S.[2], Bedford, H.[4], Cortina-Borja, M.[4], Setakis, E[5], Walton, S.[6], Fitzsimons, E.[7], Dezateux, C.[8,9], and Lyons, R.A.[3]

Submission History

Submitted:	18/07/2018
Accepted:	02/12/2018
Published:	02/04/2019

[1]Office for National Statistics
[2]Swansea University
[3]Health Data Research UK, Swansea University
[4]Institute of Child Health, University College London
[5]NHS Digital
[6]Hertfordshire County Council
[7]Centre for Longitudinal Studies, University College London
[8]Centre for Primary Care and Public Health, Barts
[9]London School of Medicine and Dentistry, Queen Mary University London
[†]Work undertaken while the author employed by Swansea University

Abstract

Background

In longitudinal health research, combining the richness of cohort data to the extensiveness of routine data opens up new possibilities, providing information not available from one data source alone. In this study, we set out to extend information from a longitudinal birth cohort study by linking to the cohort child's routine primary and secondary health care data. The resulting linked datasets will be used to examine health outcomes and patterns of health service utilisation for a set of common childhood health problems. We describe the experiences and challenges of acquiring and linking electronic health records for participants in a national longitudinal study, the UK Millennium Cohort Study (MCS).

Method

Written parental consent to link routine health data to survey responses of the MCS cohort member, mother and her partner was obtained for 90.7% of respondents when interviews took place at age seven years in the MCS. Probabilistic and deterministic linkage was used to link MCS cohort members to multiple routinely-collected health data sources in Wales and Scotland.

Results

Overall linkage rates for the consented population using country-specific health service data sources were 97.6% for Scotland and 99.9% for Wales. Linkage rates between different health data sources ranged from 65.3% to 99.6%. Issues relating to acquisition and linkage of data sources are discussed.

Conclusions

Linking longitudinal cohort participants with routine data sources is becoming increasingly popular in population data research. Our results suggest that this is a valid method to enhance information held in both sources of data.

Introduction

Background

Linking detailed routine administrative data to equally detailed survey responses can add value to population health research. Routine data can provide information on downstream services, while surveys can give insights into attitudes behind why it happened. However, the path to linked cohort and routine data is sometimes complex and complicated.

Consent for linkage must be obtained, linkage must be performed securely and accurately, and the resulting linked dataset must be made available to researchers in an anonymised format. A report commissioned by the Wellcome Trust reported all three of these issues as barriers to data linkage [1]. Routine data, received from various sources such as Electronic Health Records (EHRs), education records, or local government authorities, can create large datasets of detailed coded data. Survey-based cohort studies provide important participant-reported information, often collected at multiple points over a long period, which may not be available through routinely collected data. Linking routine data with cohort studies improves the overall detail and knowledge we have about a participant, and provides the ability to validate each data source [2], which strengthens research and reduces the knowledge gap. Such linkage combines the volume of activity data from routine data sources with rich data in cohort studies on personal circumstances, behaviours and attitudes not captured in administrative or clinical data, enhancing the value of both sources for public health research.

While many cohort studies have been collecting consent

*Corresponding Author:
Email Address: [redacted] (K Tingay)

https://doi.org/10.23889/ijpds.v4i1.579

for future linkage to multiple data sources [3-5], and some linkage has already been accomplished, the consenting and un-consenting populations may be demographically very different [6]. There are further concerns about whether broad consents obtained by cohort studies allow participants to be sufficiently informed about how their data will be used in the future [1].

Even assuming successful linkage and appropriate consent, several studies have reported significant delays in acquiring linked data from data owners [7, 2]. To a project with time-limited funding, such delays in acquiring the research data can be potentially catastrophic [8].

Linking cohort and routine data sources appears to be a useful endeavour given the ability of both to enhance each other, thus providing a richer research data source, but that this technique is not without methodological, ethical and technical challenges.

This article describes the experiences of linking a UK-wide longitudinal cohort study, the Millennium Cohort Study (MCS), to routine health records for consenting participants. The resulting linked datasets have been used to examine the health outcomes of childhood obesity, asthma, infections and injuries, and patterns of health service utilisation, including timeliness of immunisation in childhood [9, 10].

Objectives

The purpose of this paper is to describe the methodological issues, successes and challenges encountered in linking routinely collected health datasets in England, Scotland and Wales to singleton births from the MCS cohort.

In particular, the paper will discuss the following research questions:

1. What routine health data sources are available to link to the consenting MCS cohort, given the project research areas?
2. What are the linkage rates for each of these data sources?
3. Are there any demographic differences between the consented, linked cohort and the full UK MCS cohort?
4. What are the challenges in acquiring and linking cohort data to routine data?

Methods

Study Participants

The study consisted of Millennium Cohort Study participants who were interviewed in Wales or Scotland at age seven, and whose legal parent/carer had given consent to link the cohort member's health records at the age seven interview [3]. Linkage included only singleton members of the cohort.

Data sources

Millennium Cohort Study

The Millennium Cohort Study (MCS) is a prospective, longitudinal survey of children born between 2000 and 2001 in the United Kingdom (UK). The MCS is conducted by the Centre for Longitudinal Studies (CLS) at University College London (UCL) [11]. The sample was originally drawn from Child Benefit records, as these were claimed by almost all families in the UK at the time. The study used a stratified cluster sampling design, and oversampled births to families living in disadvantaged areas, from the smaller UK countries and, in England, areas with high prevalence of ethnic minorities. The initial interview, taking place at nine months of age, recruited 18,552 families comprising 18,818 children (18,296 singletons; mean age 295.5 days [12]). During the interviews, information was collected on physical and mental health of the child and of their carers and on their family's demographic and socio-economic background. Families were re-interviewed when the child was aged three, five, seven, 11, and 14 years and a further interview at age 17 began in 2018.

Consent to link other data sources One of the objectives of the MCS was to extend the survey content using other linked data sources [2]. To this end, consent to link health and other administrative records to the survey responses was requested from carers at different sweeps. At the age 7 contact (MCS4), permission was sought to link to the child's health records up to the child's 14th birthday and 90.7% of parents consented.

Wording of the consent forms used at MCS4 in relation to health record linkage was developed in consultation with the then NHS Information Authority, now NHS Digital, and approved by the Northern and North Yorkshire Research Ethics Committee (Ref: 07/MRE03/32) [13].

Previous Data Linkage At the first contact, when cohort members were approximately nine months old, parental or legal guardian consent was obtained to link MCS data and the child's National Health Service (NHS) birth record and maternity episode hospital records in all four UK countries [14, 13]. Of the 18,552 parents interviewed at MCS1, 92% of mothers provided valid consent to link their health records [3, 14, 15]. Health record linkage was restricted to data relating to the pregnancy and birth of the cohort member [13]. Birth records were obtained from the National Health Service Central Register (NHSCR). Linkage was conducted using identifying information specified by the health data controller and included combinations of the child's and mother's names and dates of birth, father's name, child's sex, birthweight, and birth order (if part of a multiple birth), and name of the hospital of birth [14]. Linkage rates for England, Wales, Scotland and Northern Ireland health authorities ranged from 83% to 92%, with the highest rates in Scotland and Wales [14].

Routinely collected health datasets

The selection of routine datasets for linkage to the MCS was based on their availability and relevance to the research questions. In line with the consent to link, linkage was conducted based on whether the cohort member appeared in the target dataset at any point before their 14th birthday (maximum September 2015). However, as these are health datasets, absence of linkage does not indicate a failure in the linkage method. Not all children attend Emergency Department or hospital inpatient settings and, therefore, not all children linked via a health spine will appear in these datasets. The

Table 1: Datasets, with respective years, requested from Scotland and Wales for linkage with Millennium Cohort Study data for consented cohort members..

Dataset setting	Wales	Scotland
Child Health (including immunisations)	2000-2015	2002-2015 (imms) 2011-2015 (CHSP)
Emergency Department	2009-2015	2007-2015
Hospital inpatient	2000-2015	2000-2015
Primary Care General Practice	2000-2015	Not available

Observation: The maximum date for which data was requested was August 31st 2015, although, for consent reasons, linkage is only conducted up to the child's 14th birthday. CHSP = Child Health Systems Programme; imms = immunization dataset.

full list of datasets with years from which data are available is given in Table 1.

Child health datasets Scotland and Wales both have unified child health datasets comprising birth, physical, developmental, and immunisation data, among other relevant information pertaining to the child's health and wellbeing. In Scotland, this is the Child Health Systems Programme (Pre School and School) (CHSP), and the Scottish Immunisation and Recall System (SIRS) datasets, and in Wales, the National Community Child Health Dataset (NCCHD). The Scottish data collection began in 1993 for some health boards, but only became available for all Scottish health boards from 2011/12 [16]. Immunisation records have been consistently recorded in Scotland since 2002.

Given that some Scottish Health Boards did not start returning child health data until 2011, the MCS cohort may not have sufficient data for some research questions concerning child health up to age 14 years. Child health data has been collected in Wales from 2000 onwards and hence covers the entire period of MCS.

Emergency department datasets Emergency Department data are collected in both Scotland and Wales. In Scotland, this is the Accident and Emergency version 2 dataset (A&E2), and in Wales the Emergency Department Data Set (EDDS). Scotland holds data from 2007, with diagnosis, injury fields, and an alcohol involvement flag added in 2010. Data collection began in major Welsh hospitals in 2009, and was extended to other emergency clinics in 2012.

Since the MCS cohort children were born in 2000 and 2001, neither emergency department datasets cover events occurring before age nine years.

Hospital inpatient datasets Scotland and Wales hold hospital inpatient and day patient datasets. In Scotland, this information is available through the General Acute Inpatient and Day Case – Scottish Morbidity Record (SMR01) dataset, and in Wales through the Patient Episode Database for Wales (PEDW). Both Scottish and Welsh hospital inpatient datasets covered the period 2000-2015.

Primary Care General Practice datasets In Wales, approximately 70-80% of General Practitioners (GPs) contribute linked data for sharing with researchers. This information is available through the Welsh Longitudinal General Practice (WLGP) dataset. When a practice signs up to share data, all the historical data are uploaded. For this study, GP records were available from 2000 to September 2015. The data include diagnoses, test results, and prescriptions issued, although the dispensing of these prescriptions is handled by separate pharmacy systems. Scotland does not yet have a national GP dataset.

Data storage environment

Data were stored and accessed in the Secure Anonymised Information Linkage (SAIL) databank held at Swansea University in Wales.

Linkage

The UK NHS is devolved across the four countries, England, Wales, Scotland and Northern Ireland, with different arrangements for supporting the provision of routinely collected health data to consented studies. Linkage is provided in Scotland by the NHS Information Standards Division (ISD), in Wales through NWIS and the SAIL Databank, and in England by NHS Digital. For the purposes of this project, data controllers in Scotland, Wales and England were approached. Northern Ireland was not included due to relatively small population size compared with the other UK countries [12].

Consent to link to health records in England was originally sought in 2009 but, due to multiple reorganisations in the NHS informatics organisation, changes in staff and differing interpretations of the wording of consent forms by individuals over time, we were unable to obtain agreement for linkage within the time frame of this study.

Anonymisation

Linkage for Scotland and Wales was approved by the appropriate data controllers and conducted by Trusted Third Party (TTP) NHS Information Services: The Scottish Information Services Division (ISD) and the NHS Wales Informatics Service (NWIS).

Linkage variables and methods

Both Wales and Scotland use a mixture of deterministic and probabilistic methods [17-19], based around a central population spine with lexical and soundex matching to account

for spelling differences. For both sites, linkage specificity is >99% and sensitivity is between 95-100%, depending on the data source [17-19].

Since 1995, The National Health Services (NHS) in Wales assign a unique number, the NHS number, to all babies. In Scotland, the same function is served by the Community Health Index (CHI) number. These unique numbers populate many of the administrative health databases in the UK.

Linkage for both sites is described in Figure 1 and is based on a combination of NHS identifier (NHS Number or CHI), all or part of the first name, surname, date of birth, address and postcode of residence.

Compared with the previous linkage to mother's maternity hospital record and cohort member's birth record [14], linkage for this project relied more on the cohort member's details and less on the mother's details.

Transfer to and linkage within SAIL

For linkage purposes, both TTPs assigned identifiers common to the MCS and routine health data sources. In Scotland, health data from cases matched on the CHI number were transferred directly to CLS with the CHI number replaced with the project-specific MCS identifier. These data were then securely transferred to the SAIL Databank and provisioned to the project. The data permissions and flow process is shown in Figure 2.

NWIS uses an encrypted Anonymous Linkage Field (ALF), based on the symmetrical block cipher Blowfish algorithm [20], for all linked datasets. Welsh health data are stored centrally under appropriate permissions and provisioned following approval by the local Information Governance Review Panel and CLS MCS application based on the ALF linkage already being completed, unless new data are needed for a project. A look-up table maps between MCS and health datasets using both ALF and MCS identifier.

Study cohort representativeness

The UK MCS cohort includes Welsh and Scottish cohorts because it was not possible to exclude these from the full analyses, given the subset of data available for this project. The Welsh study sample included here comprises 13.3% of the UK MCS cohort, and the Scottish, 10.3%. The MCS cohort is itself not fully representative of the general UK population, being over-representative of children from lower socio-economic areas [3].

Because the sub-population sizes are so different, all figures relate to percentages, rather than household size, which is reported as the mode for each sub-population. Percentages were weighted using survey and non-response weights to account for the clustered sampling, attrition between contacts, and consent to data linkage [21].

Given the subpopulations are dependent, and the data are proportional in nature, we were not able to perform statistical tests to measure the degree of difference between these populations.

Results

Acquiring the datasets

Linkage applications were approved for Wales and Scotland, but were declined for England after 2 years, owing to concerns about the wording of the consent forms. Hence, the remainder of this report focuses on experience of linkage for Scotland and Wales only.

For Scotland, the original application was submitted in March 2015, approved in July that year, with data received by CLS in May 2016. The majority of the process involved linking the cohort to their health records. Once linked data had been released to CLS, a second application was required to share the data with the project team, which took one month.

For Wales, obtaining data took approximately 3 months, as the data were already linked and held within SAIL, and work could commence processing and analysing the datasets as soon as IG approval had been obtained in April 2015.

Participants

Of the original 18,552 families interviewed at MCS1, 2,760 were interviewed in Wales, and 2,336 in Scotland. At MCS4, 1,965 families (with 1,951 singletons) were interviewed in Wales and 1,623 (1,598 singletons) in Scotland. After excluding those families who moved out of the UK, non-singleton children, those without consent to link health data, and the English and Northern Irish cohorts, the overall baseline study population was 1,838 Welsh children and 1,431 Scottish children (see Figure 3).

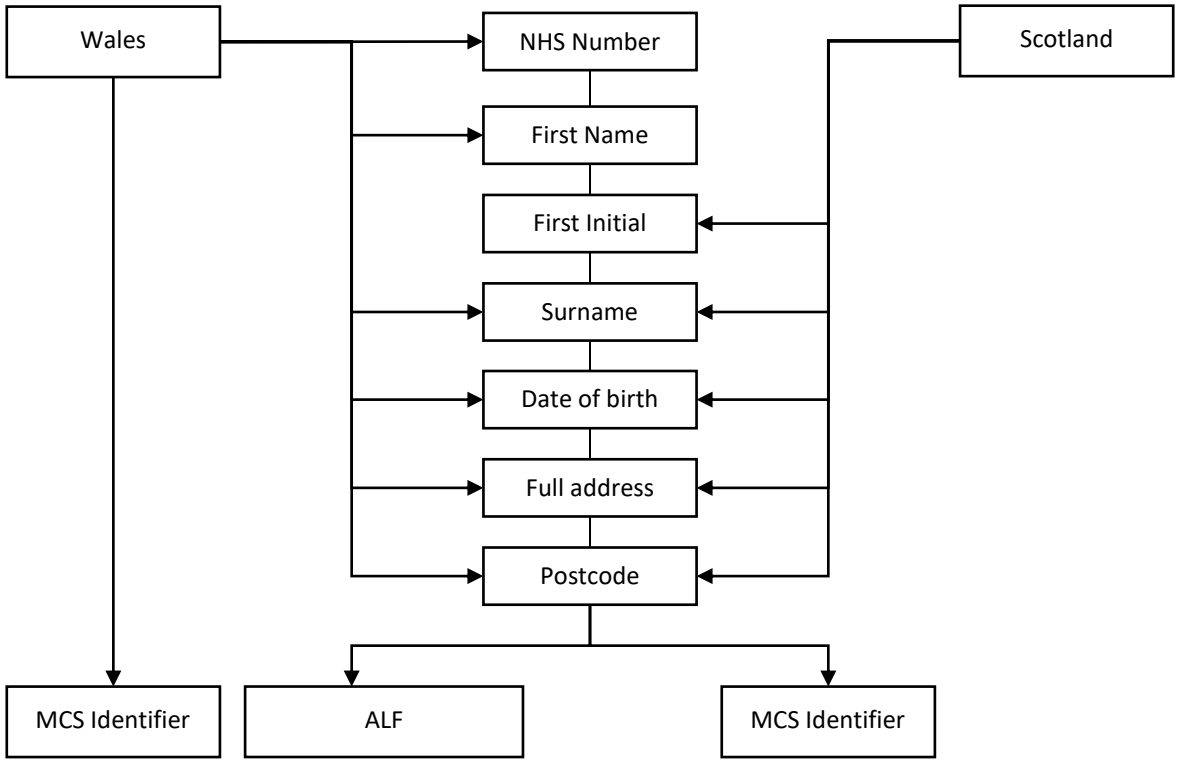
Linkage was successfully completed for almost all children for whom consent had been given. In Wales, 99.9% of children were matched to WDS and in Scotland 97.6% of those sent to ISD were matched to CHI.

Because linkage was conducted on population spines rather than for individual datasets, with unique identifiers then mapping across multiple data sources, separate specificity and sensitivity analyses were not required. We were not able to measure population-level coverage because, although we had access to data for all Welsh children, we only received Scottish data for the linked cohort.

Table 2 gives the number of children with at least one record in each dataset. For Wales, child health and GP datasets had the highest proportion of records, at 99.6% and 83.6% of the project baseline population, respectively. The Scottish child immunisations dataset contained 100% of the project population, but linkage was lower for the general Scottish child health dataset (91.3%). It appears that the Scottish immunisation dataset contains records on children who do not receive any immunisations, whereas the Scottish child health dataset only contains information on children who have interactions with the service. To avoid confusion with the Scottish child health dataset, Table 2 omits the separate Scottish immunisations dataset.

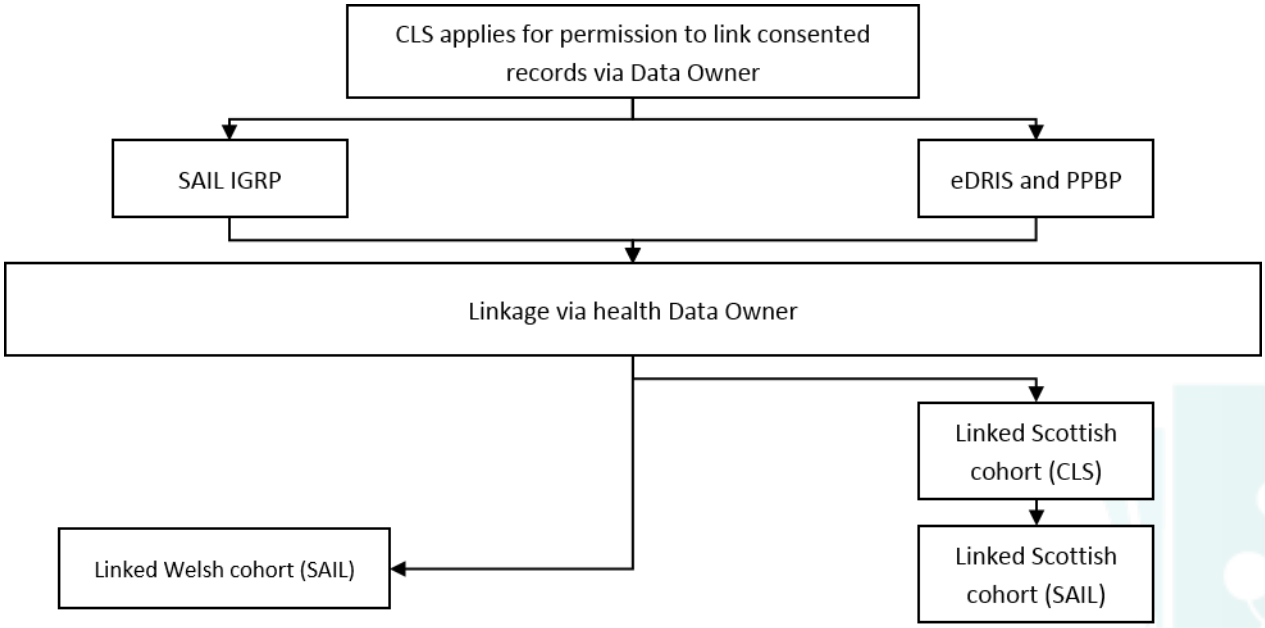
Inpatient records were present for 72.6% of the Welsh cohort, and 58.2% of the Scottish cohort. Sixty-five percent (65.3%) had attended a Welsh, and 72.2%, a Scottish, Emergency Department at some point between their 9th and 14th birthdays.

Figure 1: Information required by health service Trusted Third Parties to create linkage identifiers for this project



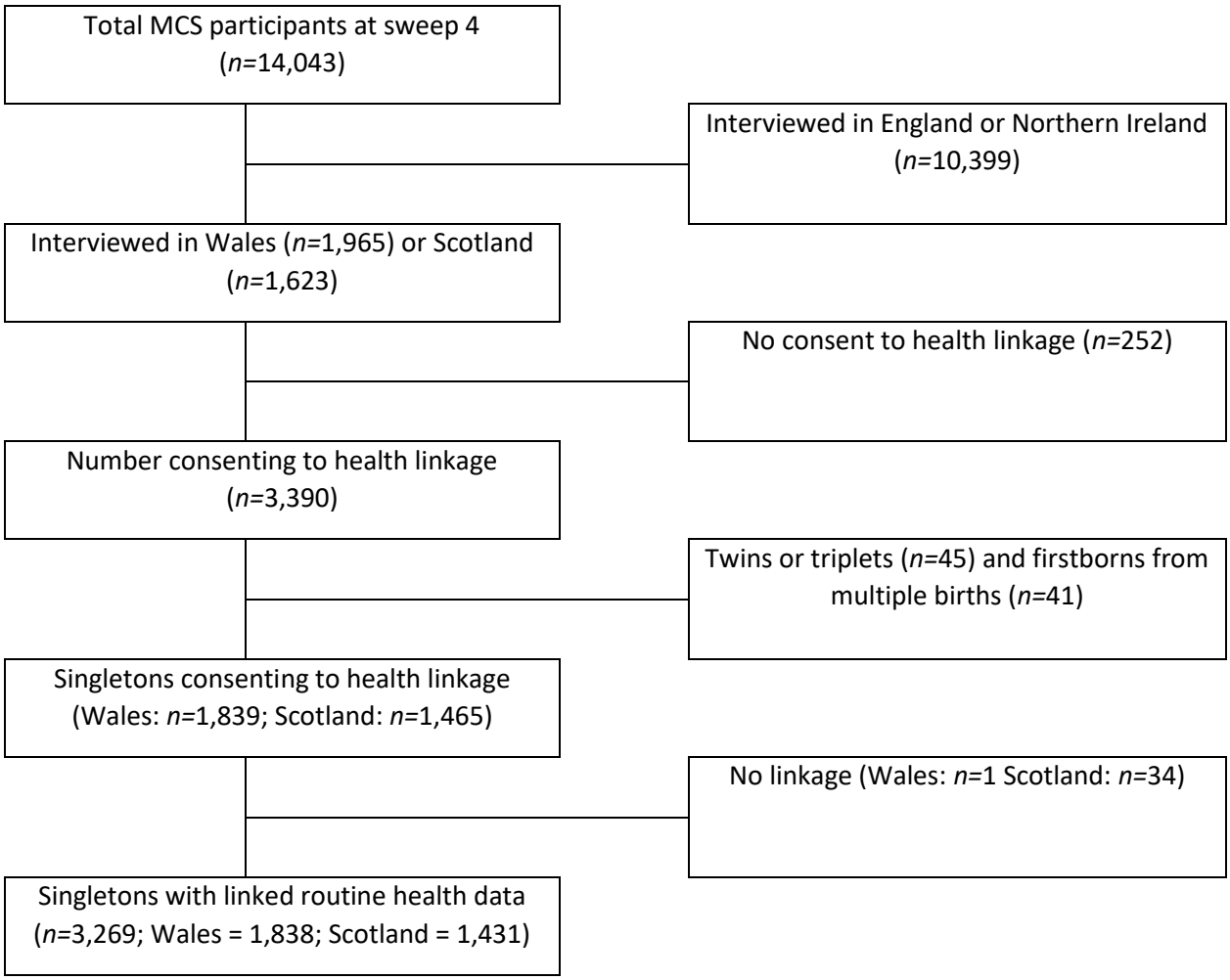
Observation: MCS Identifier is the anonymized field assigned by the Millennium Cohort Study team. ALF is the Welsh Anonymized Linkage Field assigned by the Welsh Trusted Third Party

Figure 2: Data governance and flow diagram.



Observation: CLS = the Centre for Longitudinal Studies; SAIL IGRP = the ethics board for projects wishing to use the SAIL databank and Welsh linked data; eDRIS and PPBP = ethics boards for projects wishing to use Scottish routine health data

Figure 3: Flow chart showing inclusion criteria and numbers for the final linked population



Observation: MCS = Millennium Cohort Study

Table 2: Family, maternal, and child characteristics of the sample (N=162,847) across child health groupings

Dataset setting	Wales (N = 1838)	Scotland (N = 1431)
Child Health (excluding immunisations)	1831 (99.6)	1306 (91.3)
Emergency Department	1200 (65.3)	1033 (72.2)
Hospital inpatient	1334 (72.6)	833 (58.2)
Primary Care General Practice	1537 (83.6)	Not available

Study cohort representativeness

Table 3 outlines general demographic details for the linked sub-populations compared with the full MCS cohort. Percentages for cohort member gender and main respondent at birth of cohort member are from MCS1 responses. Otherwise, the results refer to MCS4.

Overall, the Welsh and Scottish MCS populations are not noticeably different from the UK MCS cohort, nor were the linked health data populations different from the national Welsh and Scottish consenting MCS populations. The Welsh and Scottish populations had fewer cohort members from non-White ethnic groups, and the main respondent was less likely to have semi-routine or routine job and more likely to be in a "Not applicable" employment category. For socio-economic status, "Not applicable" includes unemployed, full-time students, and unstated and unclassified employment categories.

The Welsh and Scottish cohort members were similar to the UK MCS cohort in terms of gender (53%-50.2% boys, respectively, compared with 50.7% boys in the UK MCS), and the linked routine data populations were not markedly different from the national base populations. However, there were more boys linked to the emergency department and inpatient datasets for both Wales and Scotland than in the base MCS populations (56.8% and 56.3% compared with 53% for Wales, and 52.3% and 53.3% versus 50.2% for Scotland).

Scottish cohort members are more likely to live with both natural parents (70.3%) than the Welsh cohort (66.3%), but less likely than the UK MCS cohort (72.1%). Welsh cohort members are more likely to live with their natural mother only, or to live in another family type. The mode household size was similar among all sub-populations.

For all sub-populations, the main respondent was overwhelmingly the natural mother at both MCS1 and MCS4 (MCS1: 99.7%-100%; MCS4: 96.6%-100%).

Welsh main respondents were more likely to give birth to the cohort member when aged less than 30 years (56.3%) than those in the Scottish (49.8%) or UK MCS (55.4%) cohorts. Scottish main respondents were more likely to be 30 years of age or older at birth of the cohort member (50.2%) when compared to Welsh (43.6%) or UK MCS (44.7%).

All of the linked populations showed higher levels of young mothers (aged 12-29) and lower levels of older mothers (aged 30+) than the national base populations.

Discussion

Key results

Linkage of demographic details on NHS health registers to MCS cohort participants by NHS Trusted Third Party organisations in Scotland and Wales was feasible and was completed for virtually all children. While the Welsh data were obtained quickly, owing to the availability of linked data in the SAIL databank, delays in acquisition of Scottish data, and non-acquisition of English data, is consistent with the experiences of others [7, 2]. That delays in acquiring data has been reported by multiple countries suggests that this is a broader research issue requiring further guidance.

Linkage matching of health datasets was higher than the 92% for both Wales and Scotland reported in the birth registration study [14]. Due to the nature of the data received, we are not able to test completeness of linkage for these specific datasets. We only received health data for the linked Scottish cohort, for example, and not for the full Scottish or Scottish child population. We are therefore unable to measure linkage rates for individual datasets. However, as linkage was conducted on population spines for both countries, with the same unique identifiers being used to record individuals in different data sources, and linkage was high for these population spines, we propose that linkage rates are likely to represent health service use rather than linkage success. Not every child will have had a hospital inpatient or emergency department episode and so will not appear in these datasets. Population-based health service usage rates are poorly reported in the literature, making this difficult to measure. However, the health services through the NHS are free to access, meaning that the population is unlikely to have financial restrictions on healthcare [22]. A report by the Nuffield Trust found that children living in more deprived areas of England are more likely to attend hospital emergency departments and to have more preventable emergency hospitalisations (51.5% of the population) than those from less deprived background (32.6%) [23]. These findings are lower than our linkage rates, but our inpatient subpopulations were not restricted to preventable emergency admissions.

As expected, virtually all children had records in the Welsh child health dataset, as this dataset contains information pertaining to the child's birth, development and immunisations. The Scottish child immunisations dataset contains a record for each child, even those with no immunisation records, whereas the Welsh immunisation dataset only contains records where the child has been given at least one immunisation.

The lower number of retrieved records in the GP data is likely to reflect the percentage of GP practices contributing data to SAIL, and that not all healthy children will have attended a GP [24]. Most of the Welsh children had GP records (83.3%). This is higher than overall SAIL coverage of 78% and is likely due to partial coverage of records when patients move between SAIL and non-SAIL data-providing practices. Accurately describing GP coverage is a challenge as the system is dynamic, with practices being created, merged or closed. Hence, unless there is complete coverage of GP systems, the GP record will be partial for some of the cohort.

Demographically, our linked study population of consenting, linked singletons do not noticeably differ from the entire MCS population in many aspects, which suggests a relatively representative sample within the MCS cohort. This is particularly relevant for the disorder-specific aspects of our study (see, to date, [9][10]). The subpopulations for the different linked health datasets were also not markedly different from the base study population.

Our sample has a greater number of white respondents compared with the full MCS population, reflecting the sampling design and the fact that most people from black and minority ethnic groups reside in England, and has a higher proportion of people in uncategorised employment main respondents. Wales and Scotland have overall White British populations of 97.6% and 96% respectively, compared with 93.26% in England [25, 26].

Our sample also contains a higher proportion of children whose parents are not employed or who are in a "not stated" employment category. As the MCS cohort was deliberately

Table 3: Demographics for the linked and total Millennium Cohort Study populations.

	All	Wales					Scotland			
	MCS	MCS	NCCHD	GP	EDDS	PEDW	MCS	Child Health	AE2	SMR01
Number in MCS4	13857	1838	1834	1537	1200	1334	1431	1306	1033	833
Demographics and descriptives (% unless otherwise indicated)										
CM Gender (MCS1)										
Male	50.7	53.0	52.9	53.5	56.8	56.3	50.2	50.4	52.3	53.3
Female	49.3	47.0	47.1	46.5	43.2	43.7	49.8	49.6	47.7	46.7
Main respondent age at birth of CM (grouped, MCS1)										
12-19	8.6	8.8	11.3	11.7	13.5	13.2	8.5	11.9	11.7	12.9
20-29	46.8	47.5	49.5	49.4	50.1	49.0	41.3	43.2	47.0	46.1
30-39	42.5	40.7	37.6	37.6	35.4	36.5	46.3	42.7	39.0	38.9
40+	2.2	2.9	1.5	1.3	1.1	1.3	3.9	2.1	2.3	2.1
Mode of household size	4	3	3	3	3	3	3	3	3	3
Main respondent relationship to CM (MCS1)										
Natural mother	99.7	99.8	99.7	99.7	99.6	99.7	100.0	99.7	100.0	99.5
Main respondent relationship to CM (MCS4)										
Natural mother	96.6	98.0	97.8	97.7	97.2	98.0	100.0	100.0	100.0	100.0
CM ethnicity										
White	83.7	97.2	97.2	97.4	97.0	97.2	97.6	97.3	97.3	97.6
Other	16.3	2.8	2.8	2.7	3.0	2.8	2.4	2.7	2.7	2.4
Parents/Carers in the household										
Both natural parents	72.1	66.3	66.4	65.9	63.3	65.2	70.3	69.6	69.2	67.7
Natural mother and step-parent	5.0	6.8	6.8	6.9	7.0	6.7	6.8	6.6	6.9	7.9
Natural mother only	20.2	23.5	23.5	23.6	25.5	24.9	21.1	22.0	22.2	23.2
Other	2.6	3.5	3.3	3.6	4.2	3.2	1.8	1.8	1.7	1.2
SES										
Management and Professional	28.8	23.6	23.6	22.8	22.3	21.8	23.4	24.1	22.5	19.0
Intermediate	17.4	11.8	11.9	12.1	12.7	11.3	14.8	14.4	15.4	15.7
Small employer and self-employed	7.1	5.0	5.0	5.2	5.0	5.5	4.7	5.0	3.8	4.9
lower supervisory and technical	4.4	3.1	2.9	2.9	2.8	3.5	3.1	2.6	3.2	3.3
Semi-routine and routine	33.8	18.8	18.8	18.4	19.6	18.5	19.2	18.9	19.5	19.9
Not applicable	8.5	37.7	37.8	38.6	37.5	39.4	38.8	35.1	35.5	37.3

Observations: MCS = Millennium Cohort Study; NCCHD =National Community Child Health Dataset; GP = Welsh Longitudinal General Practice dataset; EDDS = Emergency Department Data Set; PEDW = Patient Episode Database for Wales; Child Health = Child Health Systems Programme; AE 2 = Accident and Emergency version 2 dataset; SMR01 = General Acute Inpatient and Day Case – Scottish Morbidity Record; CM = cohort member; SES = Socioeconomic status.

chosen to over-represent more deprived areas, care must be taken when applying research findings from this linked cohort to the wider population, especially in relation to emergency admissions [23]. Our findings of higher rates of younger mothers than the base populations may also reflect higher rates of deprivation in Wales and Scotland, as deprivation has been found to be both a factor leading to, and an outcome of, young motherhood [27].

While there were no differences between the Welsh and Scottish base MCS cohort and the linked health dataset populations, there were noted differences in gender and maternal age, especially for inpatient and emergency department datasets. Several studies have found male children to be more likely to receive healthcare than female children, especially in inpatient and emergency departments [28-30]. This disparity has been speculated to be due to both cultural, physiological and behavioural differences in gender. Piccini et al reviewed several studies that found boys appeared to have preferential access to healthcare in non-Western cultures [29]. Both Piccini et al and Hon and Nelson [28] reported differences in rates of disease among boys and girls, although it is not always the case that boys have higher rates of disease. McQuinn and Campbell found gender-related emergency department attendance to be related to the child's choice of sporting activity, with boys tending to play more contact sports than girls [30]. It is, therefore, not unsurprising that our emergency department and hospital inpatient subpopulations have higher rates of male children than female children. If anything, our findings of 53.3% and 56.3% boys for Scottish and Welsh inpatient datasets respectively are lower than Hon and Nelson's average of 59% boys. These findings will be explored further in health-specific research within this project.

Data from routine sources for consenting cohort members in a longitudinal survey were retrieved for between 70% and 99% of participants, depending on the dataset. Given that the cohort comprised children under 14 years of age, it is perhaps not surprising that linkage to datasets relating to childbirth, pregnancy, early years, and immunisations had the highest yield of retrieved records.

Several studies have reported biases in relation to survey consents, with certain demographics more likely to give consent than others [6]. Consent to linkage for MCS participants was sufficiently high to not make this an issue, although, in performing our analyses, we have used the consent weights developed by Sera et al. [21]. However, as with other studies, ours was hampered by the complexities of consent as interpreted by different data owners.

Differences between the structure and content of routine datasets from different UK countries and how these were harmonised in order to create comparable explanatory and outcome variables for research will be covered in a separate paper.

Currently, linking routinely collected data to survey data requires informed consent from participants. It can be difficult to future-proof consent forms, despite best efforts using available governmental guidance [31]. Securing enduring consent against a changing information governance landscape is challenging, as the current standards at the moment when consent is obtained may not be acceptable at later stages. Despite the rise in research using large linked datasets over the last decade, uncertainty remains regarding how to ensure adequacy of consent to link to other data sources and whether this is consistently interpreted. This is particularly apparent in the disparities we experienced in approvals to link the health data between each UK country. While this confusion may be resolved when the MCS cohort is re-consented at the next sweep, the lack of consistency between data custodians can present a significant obstacle to successful completion of funded projects. It could be argued that ignoring participants' expressed wishes to have their data linked for research would be detrimental to public perceptions of research [32]. More research is needed to determine public attitudes for consent to link routine data to survey responses, especially in the case of longitudinal child studies when children whose parents gave consent become able to consent themselves [33, 34].

This study builds on the previous MCS linkage work, both in validating linkage consistency, and, more importantly, supplementing the amount and type of linked data to this cohort. Linkage to longitudinal routine health records has the ability to provide a rich research resource for further studies. The resulting linked cohort has been used to better understand timeliness of childhood vaccination [9] and comparing GP-rated versus maternal-reported history of childhood wheezing [10].

Limitations

The amount of data expected but not acquired from NHS Digital yielded less statistical power than one corresponding to a large study population. Thus, some results are only descriptive, although they are generalisable to Wales and Scottish populations and incorporate survey weights to account for attrition and sampling design. An attempt to access linked English hospital data will be made at a later stage once agreement has been reached between CLS and NHS Digital. CLS are currently obtaining consent for linkage of health records at the age 17 MCS sweep (MCS7). Re-consenting the MCS cohort will enable longer-term follow-up.

Lack of access to the non-consenting MCS population means that we are unable to look at differences between the different groups of linked-consenting, unlinked-consenting, and unlinked singleton births. However, this could be an area of further research. The MCS team have published a report on PEDW linkage using the full consenting dataset [35], but more research is needed to compare the different groups in order to better understand the linked data as a research resource.

Similarly, while our sample is largely representative of the wider MCS cohort, and of local ethnicities, it is not clear how this linked cohort differs from the wider Welsh and Scottish population. Unfortunately, it was not possible to compare linked and unlinked populations as part of this project, but we would recommend this as a useful area of future work.

Lessons learned

Our study shows that linking national cohort responses to routine health data across multiple jurisdictions has the potential to create large and complex research datasets. However, based on our experiences, we would recommend that researchers wishing to create new datasets from previously unlinked data sources obtain approval in principle from the data owners prior to starting the project. While we would advocate for the reuse of previously-linked data sources for future research (pending approval by the data owners and ensuring

consent is respected), obtaining approval to link new data sources can be time-consuming beyond the project timeframe. Some data sources, such as NHS Digital in England, publish minutes from their ethical approval panels. We would advise researchers, in the early stages of project development, to familiarise themselves with the types of projects that have been approved by their chosen data sources. This may identify potential delays if the proposed data use has not previously been approved. It would also be advisable to have representation from data owners on the project Steering Committee.

Harmonising the health data from different countries is sufficiently detailed to be tackled in a separate article. We would, however, recommend that researchers using data sources from similar settings over multiple countries familiarise themselves with the metadata for each, and include harmonisation as a pre-analytical process in their work plan.

Conclusions

Our project has found that linking cohorts to routine health data is challenging but worthwhile, as the linkage rates are high enough to potentially provide valuable additional research information. The linkage produced for the project have already been used to measure childhood immunisations and respiratory conditions, with reports from research into injuries and physical activity in progress. The linkage enables both additional information and the opportunity to validate the different data sources against each other, thus providing both enriched data and methodological rigour.

However, there are issues around inconsistent handling of consent between data providers, and in the length of time taken to acquire the data. Until these issues are addressed, researchers should consider these potential delays when planning their projects, and data custodians could look to proactively acquiring datasets for research use.

Linking survey and routine data is a useful research tool. As with issues around consent, greater consistency between distinct but related data owners both regarding access to, and use of, the data is needed by the research community and wider public in order to make full use of the potential linked cohort and routine data can offer to researchers.

Other information

Funding

This work was supported by the Wellcome Trust (grant number 087389/B/08/Z). For this project, KST was supported by awards establishing the Administrative Data Research Centre Wales from the Economic and Social Research Council (ESRC). CD, RAL and AA are supported by awards establishing the Farr Institute of Health Informatics Research from the Medical Research Council (MRC), in partnership with Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the ESRC, the Engineering and Physical Sciences Research Council, the National Institute for Social Care and Health Research (Welsh Assembly Government), the Chief Scientist Office (Scottish Government Health Directorates) and the Wellcome Trust (MRC grants MR/ K006584/1 and MR/K006525/1, respectively). RAL is also funded by the Asthma UK Centre for Applied Research (AUK-AC-2012-01). The Millennium Cohort Study is funded by grants to the Centre for Longitudinal Studies at the Institute of Education from the Economic and Social Research Council and a consortium of government departments. The study sponsors played no part in the design, data analysis and interpretation of this study, and the writing of the article or the decision to submit the paper for publication; the authors' work was independent of their funders.

Acknowledgements

The authors are grateful to the Centre for Longitudinal Studies, UCL Institute of Education and the UK Data Service as well as the providers of anonymised data held in the Secure Anonymised Information Linkage (SAIL) system, which is part of the national e-health records research infrastructure for Wales. The co-operation of the participating families is gratefully acknowledged.

The authors wish to thank Carole Morris from eDRIS and Gareth John from the NHS Wales Informatics Service for their assistance in acquiring and linking routine health data for this project, to Jon Johnson from the CLOSER team at the Institute of Education, and to the MCS families.

Statement on conflicts of interest

None to declare.

References

1. Public Health Research Data Forum. Enabling Data Linkage to Maximise the Value of Public Health Research Data: full report. UK: Wellcome Trust; 2015.

2. Mountain, J.A., Nyaradi, A., Oddy, W.H., Glauert, R.A., de Klerk, N.H., et al. Data linkage in an established longitudinal cohort: the Western Australian Pregnancy Cohort (Raine) Study. Public Health Research & Practice, 2016:26:e2631636, https://doi.org/10.17061/phrp2631636.

3. Shepherd, P. Consent to linkage to child health data in the Millennium Cohort Study. London: Centre for Longitudinal Studies; 2013.

4. Al Baghal, T. Obtaining data linkage consent for children: factors influencing outcomes and potential biases. UK: University of Essex; 2016. Understanding Society Working Paper Series. 2016-03. https://doi.org/10.1080/13645579.2015.1064635

5. Tingay, K.S., Heaven, M. and Lowe, S. Review into the capability of the ' Secure Anonymised Information Linkage ' (SAIL) Databank to provide data for the Social Services National Outcomes Framework for people who need care and support and carers who need support. Cardiff : Welsh Government, 2015. 71/2015.

6. Sakshaug, J.W., Couper, M.P., Ofstedal, M.B., Weir, D.R. Linking survey and administrative records: Mechanisms of consent. Sociological Methods & Research. 2012:41:535-569, https://doi.org/10.1177/0049124112460381.

7. Andrew, N.E., Sundararajan, V., Thrift, A.G., Kilkenny, M.F., Katzenellenbogen, J., Flack, F., et al. Addressing the challenges of cross-jurisdictional data linkage between a national clinical quality registry and government-held health data. Australian and New Zealand Journal of Public Health. 2016:40:436-442, https://doi.org/10.1111/1753-6405.12576.

8. Dattani, N., Hardelid, P., Davey, J., Gilbert, R. Accessing electronic administrative health data for research takes time. Archives of Disease in Childhood. 2013:98:391-392, https://doi.org/10.1136/archdischild-2013-303730.

9. Walton, S., Corinta-Borja, M., Dezateux, C., Griffiths, L.J., Tingay, K., Akbari, A., et al. Measuring the timeliness of childhood vaccinations: Using cohort data and routine health records to evaluate quality of immunisation services.Vaccine. 2017:35:7166-7173, https://doi.org/10.1016/j.vaccine.2017.10.085.

10. Griffiths, L.J., Lyons, R.A., Bandyopadhyay, A., Tingay, K.S., Walton, S., Cortina-Borja, M., et al. Childhood asthma prevalence: cross-sectional record linkage study comparing parent-reported wheeze with general practitioner-recorded asthma diagnosis from primary care electronic health records in Wales. BMJ Open Respiratory Research. 2018:5(1):e000260, https://doi.org/10.1136/bmjresp-2017-000260.

11. Joshi, H., Fitzsimons, E. The UK Millennium Cohort Study: the making of a multi-purpose resource for social science and policy in the UK. Longitudinal and Life Course Studies. 2016:7:409-430, https://doi.org/10.14301/llcs.v7i4.416.

12. UK Data Service. Millennium Cohort Study. UK Data Service. [Internet] [Cited: 12 6, 2017.] Available from: nesstar.ukdataservice.ac.uk/webview/.

13. Shepherd, P. Millennium Cohort Study: Ethical review and consent. London: Centre for Longitudinal Studies; 2012.

14. Hockley, C., Quigley, M.A., Hughes, G., Calderwood, L., Joshi, H., Davidson, L.L. Linking Millennium Cohort data to birth registration and hospital episode records. Paediatric and Perinatal Epidemiology. 2007:22:99-109, https://doi.org/10.1111/j.1365-3016.2007.00902.x.

15. Tate, A.R., Calderwood, L., Dezateux, C., Joshi, H. Mother's consent to linkage of survey data with her child's birth records in a multi-ethnic national cohort study. International Journal of Epidemiology. 2006:35:294-298, https://doi.org/10.1093/ije/dyi287.

16. ISD Scotland. Child Health Systems Programme - school (CHSP School). National Data Catalogue. [Online] 2016. [Cited: 1 15, 2018.] Available from: http://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=11.

17. Ford, D.V., Jones, K.H., Verplancke, J-P., Lyons, R.A., John, G., Brown, G., et al The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Services Research. 2009:9:157, https://doi.org/10.1186/1472-6963-9-157.

18. Lyons, R.A., Jones, K.H., John, G., Brooks, C.J., Verplancke, J.P., Ford, D.V., Brown, G., Leake, K. The SAIL databank: linking multiple health and social care datasets. BMC Medical Informatics and Decision Making. 2009:9:3, https://doi.org/10.1186/1472-6947-9-3.

19. Kendrick, S. The Development of Record Linkage in Scotland: The Responsive Application of Probability Matching. 2012.

20. Scheier, B. Description of a new variable-length key, 64-bit block cipher (Blowfish). In: Anderson, R. (eds) Fast Software Encryption. FSE 1993. Lecture Notes in Computer Science, vol 809. Springer, Berlin 1994. Fast Software Encryption: Cambridge Security Workshop Proceedings. 1993. P.191-204, https://doi.org/10.1007/3-540-58108-1_24.

21. Sera, F. and Griffiths, L.J., Dezateux, C., Cortina-Borja, M. Technical report on the enhancement of Millennium Cohort Study data with linked electronic health records; derivation of consent weights. UK: University College London; 2018.

22. Leininger, L. and Levy, H. Child health and access to medical care. Future Child. 2016:25:65-90.

23. Kossarova, L., et al. Admissions of inequality: emergency hospital use for children and young people. UK: Nuffield Trust, London; 2017.

24. Dezateux, C., Griffiths, L.G., De Stavola, B.L., Akbari, A., Bandyopadhyay, A., Tingay, K.S., et al. Analysis of factors associated with changing general practice in the first 14 years of life in Wales using linked cohort and primary care records: implications for using primary care databanks for life-course research. International Journal of Population Data Science. 2018:3(2):477, https://doi.org/10.23889/ijpds.v3i2.477.

25. Office for National Statistics. Ethnicity by area and ethnic group. KS06 Ethnic Group. [Online] Available from: https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/populationestimatesbyethnicgroup/peegmethodologytcm77190625.pdf.

26. Scottish Government. Ethnic Group Demographics. Scottish Government. [Online] Available from: https://www.gov.scot/Topics/People/Equality/Equalities/DataGrid/Ethnicity/EthPopMig.

27. Humby, P. An Analysis of Under 18 Conceptions and their Links to Measures of Deprivation, England and Wales, 2008-10. UK: Office for National Statistics; 2014.

28. Hon, K-I, E. and Nelson, E.A.S. Gender disparity in paediatric hospital admissions. Annals of Academic Medicine Singapore. 2006:35:882-888.

29. Piccini, P., Montagnani, C. and de Martino, M. Gender disparity in pediatrics: a review of the current literature. Italian Journal of Pediatrics. 2018:44:1, https://doi.org/10.1186/s13052-017-0437-x.

30. McQuillan, R. and Campbell, H. Gender differences in adolescent injury characteristics: a population-based study of hospital A&E data. Public Health. 2006:120:732-741, https://doi.org/10.1016/j.puhe.2006.02.011.

31. Lightfoot, D. and Dibben, C. Approaches to linking administrative records studies and surveys - a review. UK: Administrative Data Liaison Service, University of St Andrews; 2013.

32. Jones, K.H., Laurie, G., Stevens, L. Dobbs, C., Ford, D.V., Lea, N. The other side of the coin: Harm due to the non-use of health-related data. International Journal of Medical Informatics. 2017:97:43-51, https://doi.org/10.1016/j.ijmedinf.2016.09.010.

33. Resnik, D.B. Re-consenting human subjects: ethical, legal and practical issues. Journal of Medical Ethics. 2014:35:656-657, https://doi.org/10.1136/jme.2009.030338.

34. Tingay, K.S. A multi-agency, evaluative data set for child and adolescent mental health. MPhil Thesis, University of London, UK; 2003.

35. Setakis, E., Fitzsimons, E. Millennium Cohort Study: A guide to the NHS PEDW (inpatient and day case) Linked Administrative Data Sets: ICD-10 codes in Continuous Spells. London: Centre for Longitudinal Studies, University College London; 2017.

Supplementary paper 02

International Journal of
Environmental Research and Public Health

MDPI

Article

The Effect COVID Has Had on the Wants and Needs of Children in Terms of Play: Text Mining the Qualitative Response of the Happen Primary School Survey with 20,000 Children in Wales, UK between 2016 and 2021

Michaela James [1,*], Mustafa Rasheed [1], Amrita Bandyopadhyay [1], Marianne Mannello [2], Emily Marchant [1] and Sinead Brophy [1]

[1] Data Science Building, Faculty of Medicine, Health and Life Science, Medical School, Swansea University, Swansea SA2 8PP, UK
[2] Play Wales, Park House, Greyfriers Road, Cardiff CF10 3AF, UK
* Correspondence: [redacted]

Abstract: Play is central to children's physical and social development. This study examines changes in children's response to questions on play opportunities between 2016 and 2021. Primary school children aged 8–11 in Wales participated in the HAPPEN survey between 2016 and 2021. The survey captures a range of information about children's health and wellbeing, including open-ended questions about what could make them happier. Text mining methods were used to examine how open-ended responses have changed over time in relation to play, before and, after the COVID enforced school closures. A total of 20,488 participant responses were analysed, 14,200 pre-school closures (2016 to pre-March 2020) and 6248 after initial school closures (September 2020–December 2021). Five themes were identified based on children's open-ended responses; (a) space to play (35%), (b) their recommendations on play (31%), (c) having permission to play (20%), (d) their feelings on health and wellbeing and play (10%) and (e) having time to play (4%). Despite differences due to mitigation measures, the predominant recommendation from children after COVID is that they would like more space to play (outside homes, including gardens), more time with friends and protected time to play with friends in school and at home.

Keywords: COVID; play; health; wellbeing; children

Citation: James, M.; Rasheed, M.; Bandyopadhyay, A.; Mannello, M.; Marchant, E.; Brophy, S. The Effect COVID Has Had on the Wants and Needs of Children in Terms of Play: Text Mining the Qualitative Response of the Happen Primary School Survey with 20,000 Children in Wales, UK between 2016 and 2021. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12687. https://doi.org/10.3390/ijerph191912687

Academic Editor: Paul B. Tchounwou

Received: 18 August 2022
Accepted: 2 October 2022
Published: 4 October 2022

1. Introduction

Play is central to children's physical, mental, social and emotional health and wellbeing and helps develop skills such as problem-solving, communication, fine motor skills and confidence [1–3]. There is also importance placed on the role of play in developing physical literacy. This is defined as the motivation, confidence, competence, knowledge and understanding to value and engage in physical activities [4]. Thus play equips children with the skills to sustain an active lifestyle into adolescence and adulthood [5]. The importance of play has been recognized by the United Nations Convention on the Rights of the Child (UNCRC), where play has been enshrined under Article 31 which calls for children to be able to participate fully and equally in recreation and leisure activity. It also calls for them to have a right to be heard and taken seriously on all matters affecting them (Article 12) and to gather and use public space (Article 15) [6]. This paper takes its definition of play from the United Nations Committee on the Rights of the Child's General Comment 17 on Article 31 of the United Nations Convention on the Rights of the Child [6]. It defines play as a behaviour, activity or process initiated, controlled, and structured by children, as non-compulsory, driven by intrinsic motivation, not a means to an end and that has key characteristics of fun, uncertainty, challenge, flexibility, and non-productivity.

Play during childhood has positive impacts on multiple important long-term health outcomes including increased physical activity, improving wellbeing in children, and helping to develop resilience [7]. It is also crucial and worthwhile for the enjoyment it brings to children and their families in the moment [8,9]. Play is fundamental for good health and wellbeing; for example, being physically active through play supports children physical and emotional development, contributing to their health and wellbeing [10]. When they play, children contribute to their immediate wellbeing and to their own development.

In March 2020, the World Health Organisation declared a global coronavirus pandemic (COVID-19) caused by the transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [11]. Countries worldwide, including in the United Kingdom implemented a range of public health measures and mitigation strategies to reduce transmission, including a strict 'stay at home' policy and a national lockdown running (March to June 2020 and December 2020 to January 2021). This meant the closure of many educational, workplace and retail settings with social gatherings all but banned outside of direct household members [12]. The implications of the restrictions had the potential to directly impact play opportunities for children due to the lack of outdoor access, closure of play spaces, time with friends, school closures and online learning [13–15]. However, while pandemic research has highlighted decreases in activity in children [16–18], there is also evidence to suggest that the pandemic may have been a good opportunity for children to play and be active [19,20].

Using open-ended responses from the HAPPEN-Wales survey regarding what children wanted to make them happier and healthier, this study aims to: examine how children's reports of play have changed pre and post-COVID and to provide key recommendations on COVID recovery plans for children based on the perspectives of children.

2. Materials and Methods

HAPPEN-Wales is a primary school network [21] that was established in 2015 following research with primary school headteachers who advocated for a more cohesive approach to prioritising health and wellbeing in schools. They advocated for bringing together schools and partners in health and research to provide evidence to make more targeted plans for schools based on the needs of their pupils [22,23]. To address this, HAPPEN co-produced the HAPPEN Survey, an online self-report questionnaire that was developed and designed alongside children, teachers and stakeholders in education [21]. The survey is completed by primary school-aged children (aged 8–11) in the school setting and captures a range of self-reported health behaviours including physical activity and sedentary behaviour, diet and dental health, wellbeing and mental health and the local community. Once complete, schools receive an individual report of pupils' group-level health and wellbeing data compared to national averages. This school report is aligned to the new Curriculum for Wales (due to be rolled out in 2022) [24], whereby Health and Wellbeing is one of six distinct curriculum areas, enabling schools to tailor their health and wellbeing plans and curriculum development based on pupils' needs.

2.1. Participants

Since 2016, primary schools have been invited to take part in the HAPPEN Survey throughout the academic year. Participation for both schools and children is voluntary, and they have the right to withdraw at any time. Schools are invited to share details of the survey (including study aims and a parent information sheet) among parents/guardians so that parents were given the opportunity to opt their child out from the survey. This opt-out method of recruiting participants was introduced in 2019 and aimed to ensure that a representative sample was recruited to reflect all children in Wales. Child assent is also obtained at the start of the survey.

2.2. Data Collection

As part of the HAPPEN survey, children are asked "*What could be done to make you happier and healthier in your local area?".* This is an open-ended question with the aim of understanding what children want and need from their local communities. This question has been asked since the survey's development in 2016 and has over 20,000 responses. We analysed the responses that related to play to examine how play has changed through COVID from pre-pandemic mitigation measures (2017–March 2020) to during/post-COVID mitigation measures (September 2020–December 2021). These time periods have been defined as pre and post by the nature of school closures at the time (Viner et al.). In March 2020, schools were required to close in Wales as part of transmission mitigation measures. Schools reopened in June as part of phased return before opening in September 2020. A period of school closures was then enforced in November 2020 until January 2021. For this study, pre-pandemic is the period prior to March 2020 with during/post-pandemic running from June 2020 to December 2021.

This open-ended question was only asked when children were attending school face-to-face, therefore during school closure periods we did not collect this information. The most recent version of the HAPPEN Survey can be seen as supplementary file S1 (S1). The process of data coding involved two researchers. The first researcher downloaded the raw data, cleaned the data, checked for duplicates, generated a unique participant ID number, and removed identifiable information (MJ). This process protects participants' anonymity by ensuring that the second researcher coding the responses and conducting the analysis could not identify individuals (MR).

2.3. Analysis

Initially, a random sample of 1000 responses were selected for qualitative thematic analysis to identify common themes from the open-ended responses to the HAPPEN survey. This was led by MR who identified common key words, frequently repeated words, and iterations of those words. These frequent words were sorted into a lookup dataset used as a reference for text mining. Opportunities for children to play can be supported or restricted through having time (children's 'free' time when they can become immersed in playing), space (how public space can support or constrain children's ability to play as well as access to designated spaces for play) and permission to play (children's subjective experiences of time and space, including factors such as a sense of freedom, permission, belonging, fear and harassment, as well as the increasing adult appropriation and control of play) as stressed in Wales—A Play Friendly Country statutory guidance to Local Authorities [2]. Therefore, this was used to underpin coding. For example, mentions of parks, gardens, and playgrounds (e.g., "*we could get a park nearer to our house*" or "*a bigger garden*" were coded within spaces to play). Responses were removed from analysis if they were left blank by the participant or did not discuss play (n = 4217).

These initial codes were used to identify themes using text mining methods (see details below) on the whole dataset of responses. The text mining involved three stages [25,26]. Stage 1, or the pre-processing stage, included tokenization (breaking the text into tokens or small sentences such as words), removing stop words and, lemmatization (identifying the base form of the word, e.g., good is the base form of better). Stage 2 involves a frequency analysis of words (e.g., in our dataset the word park is mentioned >1000 times, friends is mentioned >1000, litter is mentioned >400 and the word combination feeling safe is mentioned >300 times) followed by manual review, with an expert review of the words to ensure none are missed and only relevant words are considered. Stage 3 involves co-occurrence analysis to identify words that are associated together (e.g., play & park), identify word pairs that should be coded together, synonyms (e.g., park and parc should be coded park), link to the data and code all the pair words.

The quality of coding by the automated text mining method was compared by two researchers (MJ & MR) and the text mining method was modified to improve accuracy. Following this second round of coding, key themes were identified which include (i) Time

to play, (ii) Space to play (access to play, safety when playing, having space to play, sustainability) (iii) Permission to play (being allowed to play, having relationships that permit play), (iv) Recommendations (specific play/activity recommendations) and, (v) Health and Wellbeing and play (how play makes children happier, how play can be helped with healthy diets). These themes were then stratified by the period of pre-COVID and post-COVID.

3. Results

A total of 20,488 participants completed the HAPPEN Survey between September 2016 and December 2021 (n = 46% boys, average age = 9.97). From these responses, a total of 16,271 responses were coded which discussed play. Of this number, 12,529 responses were from 2016–March 2020 (pre-pandemic, n = 23%) with the rest from September 2020–December 2021. Across both time points, the codes showed that children discussed; space to play (35%), their recommendations on play (31%), having permission to play (20%), their feelings on health and wellbeing and play (10%) and having time to play (4%). Figure 1 shows the most frequent coded words in a word cloud.

Figure 1. Frequently coded words.

Table 1 shows a breakdown of this stratified by pre and during/post-COVID mitigation measures, with no significant difference in the priorities of children between time points.

Table 1. Codes.

Codes	Pre-COVID	During/Post-COVID
Space	35%	35%
Permission	20%	20%
Time	4%	5%
Recommendations	32%	29%
Health and Wellbeing	9%	11%

From these initial codes, additional sub-themes were identified; (i) access to spaces to play, (ii) perceived safety of local spaces, (iii) the need to improve existing spaces, (iv) cleaner spaces, (v) being allowed to play, (vi) relationships that facilitate play, (vii) health and wellbeing implications of play and, (viii) having more time to play. This can be seen in Table 2 and again, highlights (aside from the emergence of specific COVID responses) there were no significant differences in their occurrence pre and during-COVID.

Table 2. Themes.

Themes	Pre-COVID	During/Post-COVID
Access to Spaces	5%	5%
Safety	7%	8%
Improve Existing Spaces	18%	19%
Cleaner Spaces	5%	3%
Being Allowed to Play	4%	3%
Relationships that Permit	16%	17%
HWB	10%	8%
More time	3%	5%
Specific activities	32%	29%
COVID	0%	3%

A breakdown of key themes and quotes can be seen as supplementary file S2 (S2).

3.1. Specific Recommendations for Play

Throughout the responses, there were many recommendations made by children for play and sport-based activities which were grouped under the specific recommendations theme to help provide clear examples of the diverse range of activities suggested. This reflects the range in children's wants and needs from play. Some common suggestions for activities were swimming, basketball, and football. Although lots of children asked for the opportunity to do "different" activities:

"Go swimming more" (2016/2017)

"Swings in the park, better football pitch, somewhere to play basketball" (2019/2020)

"To run with my friends more and play different activities." (2020/2021)

The school setting was cited as important for this with some children saying they wanted *"more PE"* or *"more after school sports"*. This theme suggests that time allocated for school-based activity needs to be dedicated to a broad range of activities. The recommendation from this theme is to consult children (e.g., through pupil voice groups at school, community groups outside school) to identify specific wants and needs from activities and play as this would be different in different settings. For content, Figure 1 highlights how prominent the school setting is within responses and how frequently mentioned sport, exercise, equipment, and more specific forms of activities are mentioned.

3.2. Space to Play

Space to play was a significant code across all academic years and pre/post-pandemic. Much of the discussion revolved around the theme of improving existing spaces (18% and 19% respectively), and improving play equipment, fixing damage and cleaning local parks:

"Make sure that parks and other places have safe equipment" (2018/2019)

"Fix damage to parks by me" (2021/2022)

"Have a litter pick up once a week . . . " (2017/2018)

This suggests that what currently is on offer does not meet the standard of children's wants and needs and to enable them to play more, parks need to be improved. This shows that despite infrastructure in place, more needs to be done to ensure that it is actually usable. Figure 1 Shows that parks, halls, environments (generally speaking), gyms and the indoors are mentioned often as spaces to play.

Within this code, safety of spaces was highlighted. Children consistently mentioned safety concerns and fears in some of the places they would like to play including fears of illegal behaviours (e.g., drug use) as well as recommending safety equipment checks. This represents children's awareness and fears regarding illegal behaviours in their local area, either through direct observation or indirectly by adults or observed directly by them.

"There are unfriendly people hanging around my area doing drugs and smoking" (2018/2019)

"More safer and clean areas for children to play and feel more comfortable!" (2021/2022)

Some children also noted the presence of bullying and "unkind" behaviour which deters them from playing. The word 'nasty' is mentioned in Figure 1. Concerns over safety also highlighted the presence of cars:

"My road is very busy so we could get some more traffic lights." (2016/2017)

"Put speed limits on the roads" (2016/2017)

Children even made clear suggestions about how to improve the safety of their local areas with the mention of traffic flow measures. It is evident they are aware that local communities prioritise the use of cars over the safety of pedestrians. They also note the presence of litter, including *"dog poo"* which appears to be a clear deterrent to play.

"Litter and dog poo and speeding" (2016/2017)

"Put more bins so that there is less litter." (2017/2018)

Access was mentioned in terms of linking homes with facilities better including better infrastructure for children to be able to get to spaces.

"We could get a park nearer to our house." (2016/2017)

"Easier to access outdoor places" (2021/2022)

It is interesting that a significant proportion of responses discuss limited accessibility to outdoor spaces. Responses to the general spaces code highlight possible reasons including poor equipment, fears over safety and cleanliness. Interestingly, in 2021 more responses mentioned their garden as a space to play (e.g., "A bigger Garden.") which would have been in line with the emergence of COVID and subsequent restrictions seeing children needing to access play in their household. It is worth considering that some children will not have had a garden and therefore, may have had no space to play at all.

The key recommendation from this theme is to not only provide outdoor spaces that children can easily access play in local communities but also to provide spaces to listen to the concerns of children and facilitate ways in which these fears can be reduced.

3.3. Permission to Play

Permission to play was centered around relationships that permit play and how relationships with friends and family are conducive to giving implicit/explicit permission for children to play. In terms of family, the presence of parents was key. Children note how parents living separately and the business of parents is not conducive to their play. This highlights how parental figures can be role models and leaders for play in certain spaces, their time gives permission for children to play:

"Do more sport with my mum" (2016/2017)

"Mum and dad to live with each other" (2018/2019)

"To have my mum not be so busy" (2021/2022)

Having friends to play with was considered essential throughout the years including having more live nearer, "being around" friends more and having opportunities to go out with them. Under friendship, bullying was consistently mentioned as something that deterred play with the word "kind" mentioned frequently in reference to friendships and bullying behaviour:

"Stop bullying" (2019/2020)

"Be kind to people and make others to be kind to others" (2021/2022)

This also has parallels to safety concerns in the space code. Under this theme, it is evident that social, supportive and "kind" environments give permission to play. Therefore, the key recommendation from this theme is to facilitate opportunities for children to be with their friends and family where possible.

3.4. Health and Wellbeing Outcomes

The theme of health and wellbeing outcomes relating to play emerged in the academic year 2018/2019 which is in line with the announcement of the new Curriculum for Wales in which health and wellbeing is one of six distinct areas. This suggests that children had a greater awareness of health and wellbeing after this year, particularly how this would impact and be impacted by their play:

"eat more veg and stay fit and healthy (ALWAYS)" (2019/2020)

"Keep fit always get fresh air" (2020/2021)

This shows that children not only want to play, but acknowledge the benefits of being able to play and the benefits that behaviours such as healthy eating could have for play. It is also interesting that some children expressed a need to socialise and awareness of the importance of this to them. Some of the most frequently mentioned words in terms of health and wellbeing can be seen in Figure 1, particularly fruit, vegetables and healthy.

3.5. Time to Play

The theme of time to play centered around children wanting more time to play and, in terms of more time to play, this was time outside playing or more opportunities to play/be active. As with all themes, there was no significant difference between pre and post pandemic responses, just that more time was needed.

"Spending more time swimming and going outside" (2016/2017)

"Allowed to play out all the time because we like playing out" (2019/2020)

"To have more time to play" (2021/2022)

While there were implications that this meant more time to play at home, there were explicit mentions of how the school setting mitigated time to play. In particular, there were many mentions of more break times and afternoon play as well as the mention of homework being done in free time. This theme is also the only theme that mentions how time to play impacted genders differently, with girls saying there was a lack of time dedicated to activities that they liked:

"Not just boy sports at play times so girls can play as well." (2018/2019)

While device use was mentioned throughout the academic years, it becomes particularly prominent in post-pandemic. This could be because of online learning during school closures and the use of electronic equipment required. The key recommendation from this theme is to protect play time and where possible facilitate it as much as possible. This is very relevant in the school setting where break times are often used as a behaviour management technique or where afternoon breaks have been removed in favour for increased learning time.

3.6. The Impact of COVID

In 2020, children started discussing COVID-19 and the impact this was having on their ability to play and be happy. For the most part, this revolved around "stopping" transmission where children acknowledged that this needed to be done. However, when taken in conjunction with other responses:

"Get rid of COVID rules" (2021/2022)

"seeing each other a lot more but because of COVID we have not been doing that" (2021/2022)

"Easier interaction not staying away from each other (COVID)" (2021/2022)

It is evident that stopping the transmission equated to the end of the restrictions for children. These restrictions impacted socialisation and children specified that they wanted to see their friends again and not have to socially distance. It is interesting to note that children did not discuss COVID-19 in relation to the impact on their health showing that this age group was not concerned about contracting the virus. They were more concerned about the impact this was having on the ability to see their friends. The recommendation from this theme is to acknowledge that children have missed key socialisation time and to facilitate this and nurture this in recovery plans. Play is an important place for this to happen.

4. Discussion

This study aimed to examine how children's reports of play have changed pre- and post-COVID and to provide key recommendations that can inform COVID recovery plans based on the perspectives of children. There is a well-established body of solid evidence that shows the contribution that play, particularly self-organised play, can make to children's immediate and long-term wellbeing, to their physical health, problem-solving, communication, fine motor skills, confidence and to their mental health and resilience [7,27]. Therefore, it is vital we protect play and listen to the best ways in which to facilitate it.

Ten key themes were identified under the codes of Time (time to play), Space (access to play, safety when playing, having space to play, sustainability), Permission (permission to play, relationships and play), Recommendations (specific play/activity recommendations), Health and Wellbeing Outcomes (health and wellbeing and play) as a result of analysis of children's responses from 2016 to 2021. The text mining analysis showed that the presence of these themes was similar pre- and post-pandemic, showing that the recommendations for children's play remain similar throughout time and circumstance.

The specific suggestions made by children were diverse and broad, encompassing a range of answers from football, basketball, and swimming to simply asking for a variety of more activities. This highlights that there is not a one size fits all model to promote activity and play in children and therefore, it is a worthwhile pursuit to engage and involve children themselves [28]. Previous research has highlighted that older children have requested more choice in activities, rather than prescriptive, adult-led programs [28]. This study shows that a similar request is being made by younger children with a wide-range of play-based and sport-based suggestions being bought forward. Therefore, we cannot assume a 'one size fits all' model should work with play. Other studies have acknowledged that choice, or lack of, is a reason why young people become disengaged with activity, particularly girls [29]. It is important that suggestions are acknowledged and the diverse and broad range of interests and needs are considered. This can be done so via a period of consultation with children. By asking them, interventions and programs can implement exactly what they want and need, helping to improve the success, longevity, and sustainability of play-work in communities and opportunities to play in schools.

Speaking to children about their solutions could have a significant impact on designing and implementing pandemic recovery plans and is reflected across the themes emerging in this study. While it is not always feasible to provide a wide-range of opportunities due to lack of resources (e.g., space, time or funding) [29], research highlights that providing autonomy and creating partnerships between users (in this instance, children) and those delivering (e.g., teachers, play workers) could improve the sustainability and efficiency of play/activity opportunities to improve health and wellbeing [30,31]. Particularly, in the wake of the pandemic where play was reduced due to mitigation measures, it is important we look to facilitate the best opportunities for children to overcome this lost time. To do this, we need to look at their suggestions and what previous research has acknowledged as positive play/activity-enablers.

Throughout the themes children recommend protecting spaces to play with their friends as a priority. This remained a constant throughout the time periods. Space to play was the most frequently discussed amongst children, encompassing access, safety when playing and sustainability of play spaces. This was centered around local park spaces, their upkeep and maintenance, safety, and the general accessibility to these spaces. Parks have been acknowledged by previous research to be important spaces for play [32–34]. Playing in these environments supports children to feel part of their neighbourhoods and wider communities, allowing children to learn about the world around them, make connections, and develop a sense of identity and belonging [1]. Thus, this access is integral for their development and lived experience (or enjoyment). Access could mean closer proximity [34] but it could also refer to access to features within the park such as open spaces, courts or trees to climb [32]. Fears over safety may reduce access [35,36] due to concerns over encountering illegal behaviours and unsafe equipment/environments as mentioned by children in this study. Independent mobility in children and the ability to explore, play and be active in local outdoor spaces without adult supervision has declined [37,38]. This needs to be protected and work needs to be done to improve mobility for children in their local neighbourhoods.

It is vital that we listen to the concerns of children and facilitate ways in which these fears can be reduced, and access improved. Previous research has shown that perceptions of safety are significant in facilitating play/physical activity, particularly for those more deprived [20,35]. Concerns over safety were more frequently mentioned post-2019 where, in 2020, children would have been subject to strict lockdown restrictions [12]. Studies suggest that being confined to local areas during periods of restriction may improve perceptions of safety [20].

In light of COVID-19, it is vitally important that we give children space to play due to the lack of access to outdoor spaces, time with friends and school closures [13–15] that they have faced as a result of mitigation measures. During the pandemic and its associated lockdowns, access to outdoor play was particularly important as the population was confined to their homes. Outdoor play gives sense of freedom and control that children can enjoy and it allows for children to be energetic and physically active. Play is a form of physical exercise for children. However, to stop transmission, playgrounds were closed during lockdown. In Wales, as more was learned more about the pandemic and ways to manage transmission, Welsh Government advice was updated to establish the important role that playing has in supporting wellbeing. The government provided guidance and prioritised the opening of parks, playgrounds and childcare and play work provision from summer 2019 onwards, highlighting the need to play. This need remains, and as children have stated, the protection and maintenance of play spaces should still be high on the agenda. A recent study of children's experiences of playgrounds in the pandemic highlights this further with children expressing their pleasure at being back in the playground with no social distancing [39].

Alongside space, children discuss the role of relationships in supporting play opportunities. Play supports socialisation. As mentioned above, children have welcomed the removal of social distancing in facilitating socialisation [39]. Our study highlights children value the characteristics of kindness in their peers and concerns about bullying were mentioned suggesting children seek nurturing and supportive structures which will facilitate their play. In 2019, Welsh Government launched new guidance on how to stop bullying in schools [40]. Its consistent mention suggest work still needs to be to eradicate bullying from the school setting. Discussing and calling out bullying can support children to develop confidence to play interdependently outside of school.

With family members, it was apparent that children see parents/caregivers as role models and supporters for play. This is important to note as previous research has highlighted the novelty of the pandemic enabling some children to spend extra time with family members that they otherwise would not have had due to an increased number of parents working from home or being furloughed [20]. This has also been acknowledged by

school staff who reported children having more opportunities for walking and, spending time outside, with this contributing to strengthened family relationships [18]. This may then have had a positive impact on some children during this time and this time should be valued.

The pandemic saw children access their local community for exercise and play during the pandemic, mostly being accompanied by parents [41]. Care must be taken to ensure that, as restrictions on outdoor spaces is relaxed, that children feel connected to their communities to help them to gain confidence to play out. For children, their main concern with the pandemic were restrictions on socialisation and having to distance from their friends. Research has shown that these measures will have had an impact on children's wellbeing [42] with the removal of positive interactions with peers, teachers, coaches and wider family members. Therefore, as part of COVID-19 recovery plans, it is essential we value that children need this socialisation time back. With the uncertainty caused by the pandemic, opportunities to play are vital to helping children make sense of their experiences, problem-solve, reconnect with their peers, and promote their own wellbeing. As we develop interventions and initiatives to support children emerging from the pandemic and its related restrictions, play is one of the most important areas of focus to promote children's health and wellbeing.

The wider perception and impact of lockdown for children is mixed in the literature. There is research to suggest that lockdown may have been a positive experience for young people, with physical activity improving, as well as sleep and overall wellbeing for some children (those less deprived) [20]. Adult perceptions suggest physical activity decreased, with wellbeing also at risk during this time [18]. Therefore, there is some contention on what lockdown has meant for children. Yet, from a children's perspective, play recommendations remain the same with them asking for more variety of activities, more space, more time and more opportunities. COVID specific responses from children show that they were aware of mitigation measures and the impact these were having on their play and socialisation. Despite being deemed the population least vulnerable to direct harms from COVID infection, the government enforced restrictions impacted the lives of children in an unprecedented way [20]. Many restrictions were implemented without consultation with children or young people, and it is still unclear what the longer-term impacts of measures such as school closures have been on children. Children were specifically concerned about the impact this was having on the ability to see their friends. Protecting play, socialisation, and opportunities to be active is paramount in recovery plans and has been reflected in other studies in this field [18,20].

The above recommendations have come as a result of analysing children's responses and are therefore advocated and suggested by children themselves. It is evident that protecting spaces to play (including investment in maintenance, upkeep, and safety) and, facilitating opportunities for children to be with their friends are important to children to help them play. These have remained constant themes throughout time and therefore, it is evident that while these recommendations are not new learning, more needs to be done to influence policy, decision-making and funding into putting these into practice.

Limitations

This study encompasses responses from 2016 to December 2021. HAPPEN rolled-out to wider Wales in 2018 however, we cannot ensure a fully representative sample of children has been recruited across Wales. While the sampling strategy was the same for 2018–2020, data were sampled more purposefully in earlier years from South Wales which may have an influence on findings from this year. Although all schools in Wales were contacted with details regarding the HAPPEN Survey, the findings in this study only represent the views of children who took part. The perspectives captured in this study may not account for the full breadth of lived experiences of all children.

5. Conclusions

Drawing upon children's responses, it is essential that we advocate for the wants and needs of children particularly in relevance to giving children broad opportunities to play and be active, space to play (particularly in reference to safe and accessible spaces designed with children in mind), facilitating socialisation (especially in light of social distancing measures), and acknowledging how beneficial and integral play is the development, health, and wellbeing of children. We cannot overlook the importance of play. It has and always will be important to protect play. Given the importance of play for children at times of crisis, recognising this during and while emerging from a global pandemic could be an enormous step forward in terms of protecting the mental health and wellbeing of our children, and of future generations.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijerph191912687/s1. S1: The HAPPEN Survey, S2: Table S1.

Author Contributions: Conceptualisation, S.B. and M.J.; methodology, S.B., M.J. and A.B.; analysis: M.R., M.J. and A.B.; writing and presentation, M.J., M.R., E.M., A.B., M.M. and S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Centre for Population Health and Wellbeing, ADR UK and Play Wales.

Institutional Review Board Statement: The HAPPEN Survey was granted ethical approval by Swansea University's Medical School in September 2017 (reference: 2017-0033).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data are available to the research team according to ethical approval. The corresponding author is happy to provide data if required for scrutiny.

Acknowledgments: The research team would like to thank all pupils, teachers and schools who took part in facilitating, administering, and taking part in the HAPPEN Survey and for their support of the HAPPEN Network.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Gleave, J. Community play: A literature review. *Play Engl.* **2010**. Available online: https://www.playday.org.uk/wp-content/uploads/2015/11/Community-play-a-literature-review.pdf (accessed on 3 June 2022).
2. Welsh Government. *Wales: A Play Friendly Country*; Welsh Government: Cardiff, UK, 2014.
3. Dallimore, D. I Learn New Things and Climb Trees' What Children Say about Play in Wales. 2019. Available online: https://issuu.com/playwales/docs/psa_children_s_report_ (accessed on 3 June 2022).
4. Whitehead, M. The Definition of Physical Literacy. 2016. Available online: https://www.physical-literacy.org.uk/defining-physical-literacy/ (accessed on 17 January 2020).
5. Holt, N. *Positive Youth Development*, 2nd ed.; Routledge: London, UK, 2016.
6. Unicef. The United Nations Convention on the Rights of the Child (UNCRC). Unicef. 2019. Available online: https://downloads.unicef.org.uk/wp-content/uploads/2016/08/unicef-convention-rights-child-uncrc.pdf?_ga=2.110906055.398902239.1593171870-785006455.1593171870 (accessed on 3 June 2022).
7. Lester, S.; Russell, W. Play for a Change This Briefing Gives a Summary of the Key Findings of Play for a Change, a Review of Perspectives on Play, Policy and Practice Carried out for Play England by 2008. October 2007. Available online: https://www.academia.edu/415471/Lester_S_and_Russell_W_2008_Play_for_a_Change_Play_Policy_and_Practice_A_review_of_contemporary_perspectives_London_National_Children_s_Bureau (accessed on 3 June 2022).
8. Lester, S.; Russell, W. Children's Right to Play. In *SAGE Handbook of Play and Learning in Early Childhood*; Sage Publications Ltd.: Thousand Oaks, CA, USA, 2014.
9. Gleave, J.; Cole-Hamilton, I. A literature review on the effects of a lack of play on children's lives. *Play Engl.* **2012**, 34. Available online: https://www.eerg.org.au/images/PDF/A-world-without-play-literature-review-2012.pdf (accessed on 3 June 2022).
10. Whitebread, D. Free play and children's mental health. *Lancet Child Adolesc. Health* **2017**, *1*, 167–169. [CrossRef]

11. World Health Organisation (WHO). WHO Announces COVID-19 Outbreak a Pandemic. 2020. Available online: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic (accessed on 29 April 2021).
12. Mamot, M.; Allen, J.; Goldblatt, P.; Herd, E.; Morrison, J. *Build back Fairer: The COVID-19 Marmot Review. The Pandemic, Socioeconomic and Health Inequalities in England*; London Institute Health Equity: London, UK, 2020.
13. National Lottery Community Fund. *2021: Importance of Communities Set to Remain High as People Identify Loneliness and Isolation as a Key Issue to Tackle in Their Local Area*; National Lottery Community Fund: London, UK, 2021.
14. Engzell, P.; Frey, A.; Verhagen, M.D. Learning loss due to school closures during the COVID-19 pandemic. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2022376118. [CrossRef]
15. Shah, K.; Mann, S.; Singh, R.; Bangar, R.; Kulkarni, R. Impact of COVID-19 on the Mental Health of Children and Adolescents. *Cureus* **2020**, *12*, e10051. [CrossRef]
16. Wang, G.; Zhang, Y.; Zhao, J.; Zhang, J.; Jiang, F. Mitigate the effects of home confinement on children during the COVID-19 outbreak. *Lancet* **2020**, *395*, 945–947. [CrossRef]
17. Rundle, A.G.; Park, Y.; Herbstman, J.B.; Kinsey, E.W.; Wang, Y.C. COVID-19–Related School Closings and Risk of Weight Gain Among Children. *Obesity* **2020**, *28*, 1008–1009.
18. Marchant, E.; Todd, C.; James, M.; Crick, T.; Dwyer, R.; Brophy, S. Primary school staff reflections on school closures due to COVID-19 and recommendations for the future: A national qualitative survey. *medRxiv* **2020**.
19. Children's Commissioner for Wales. Coronavirus and Me. 2020. Available online: https://www.mind.org.uk/information-support/legal-rights/coronavirus-and-your-rights/coronavirus-and-sectioning/ (accessed on 3 June 2022).
20. James, M.; Marchant, E.; Defeyter, M.A.; Woodside, J.V.; Brophy, S. Impact of School Closures on the Health and Well-Being of Primary School Children in Wales UK; A Routine Data Linkage Study Using the HAPPEN Survey (2018–2020). *SSRN Electron J.* **2021**, *1*, e051574.
21. The HAPPEN Network Wales. HAPPEN Wales. 2020. Available online: www.happen-wales.co.uk/ (accessed on 3 June 2022).
22. Todd, C.; Christian, D.; Davies, H.; Rance, J.; Stratton, G.; Rapport, F.; Brophy, S. Headteachers' prior beliefs on child health and their engagement in school based health interventions: A qualitative study. *BMC Res. Notes* **2015**, *8*, 1–10. [CrossRef]
23. Christian, D.; Todd, C.; Davies, H.; Rance, J.; Stratton, G.; Rapport, F. Community led active schools programme (CLASP) exploring the implementation of health interventions in primary schools: Headteachers' perspectives. *BMC Public Health* **2015**, *15*, 238. [CrossRef] [PubMed]
24. Welsh Government. *Our National Mission: A Transformational Curriculum*; Welsh Government: Cardiff, UK, 2019; Available online: https://gov.wales/sites/default/files/consultations/2019-02/consultation-document-transformational-curriculum-v2.pdf (accessed on 3 June 2022).
25. Vijayarani, S.; Ilamathi, J. Preprocessing Techniques for Text Mining Preprocessing Techniques for Text Mining. *Int. J. Comput. Sci. Commun. Netw.* **2014**, *5*, 7–16.
26. Silge, J.; Robinson, D. *Text Mining with R: A Tidy Approach*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017.
27. Panksepp, J. Play, ADHD, and the Construction of the Social Brain: Should the First Class Each Day Be Recess? *Am. J. Play* **2008**, *1*, 55–79. Available online: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Play,+ADHD,+and+the+Construction+of+the+Social+Brain:+Should+the+First+Class+Each+Day+Be+Recess?#0 (accessed on 3 June 2022).
28. James, M.; Todd, C.; Scott, S.; Stratton, G.; McCoubrey, S.; Christian, D.; Halcox, J.; Audrey, S.; Ellins, E.; Anderson, S.; et al. Teenage recommendations to improve physical activity for their age group: A qualitative study. *BMC Public Health* **2018**, *18*, 372. [CrossRef]
29. Mitchell, F.; Gray, S.; Inchley, J. 'This choice thing really works . . . ' Changes in experiences and engagement of adolescent girls in physical education classes, during a school-based physical activity programme. *Phys. Educ. Sport Pedagog.* **2015**, *20*, 593–611. [CrossRef]
30. James, M.; Christian, D.; Scott, S.; Todd, C.; Stratton, G.; Demmler, J.; McCoubrey, S.; Halcox, J.; Audrey, S.; A Ellins, E.; et al. What works best when implementing a physical activity intervention for teenagers? Reflections from the ACTIVE Project: A qualitative study. *BMJ Open* **2019**, *9*, e025618. [CrossRef] [PubMed]
31. Theobald, S.; Brandes, N.; Gyapong, M.; El-Saharty, S.; Proctor, E.; Diaz, T.; Wanji, S.; Elloker, S.; Raven, J.; Elsey, H.; et al. Implementation research: New imperatives and opportunities in global health. *Lancet* **2018**, *392*, 2214–2228. [CrossRef]
32. Veitch, J.; Flowers, E.; Ball, K.; Deforche, B.; Timperio, A. Exploring children's views on important park features: A qualitative study using walk-along interviews. *Int. J. Environ. Res. Public Health* **2020**, *17*, 4625. [CrossRef]
33. Brown, G.; Schebella, M.F.; Weber, D. Using participatory GIS to measure physical activity and urban park benefits. *Landsc. Urban Plan.* **2014**, *121*, 34–44. [CrossRef]
34. Roemmich, J.N.; Epstein, L.H.; Raja, S.; Yin, L.; Robinson, J.; Winiewicz, D. Association of access to parks and recreational facilities with the physical activity of young children. *Prev. Med.* **2006**, *43*, 437–441. [CrossRef]
35. Report, F. Strategic Review of Health Inequalities in England Post-2010 Task Group 4: The Built Environment and Health Inequalities | Obesity Hub. 2010. Available online: http://obesity.thehealthwell.info/search-results/strategic-review-health-inequalities-england-post-2010-task-group-4-built-environment (accessed on 3 June 2022).
36. Kemple, K.M.; Oh, J.H.; Kenney, E.; Smith-Bonahue, T. The Power of Outdoor Play and Play in Natural Environments. *Child Educ.* **2016**, *92*, 446–454. [CrossRef]

37. Marzi, I.; Reimers, A.K. Children's independent mobility: Current knowledge, future directions, and public health implications. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2441. [CrossRef]
38. Loebach, J.; Gilliland, J. Neighbourhood play on the endangered list: Examining patterns in children's local activity and mobility using GPS monitoring and qualitative GIS. *Child Geogr.* **2016**, *14*, 573–589. [CrossRef]
39. King, P.; Gregory, C. Children's views on social distancing and playing on an adventure playground. *J. Childhood Educ. Soc.* **2022**, *3*, 48–59. [CrossRef]
40. Welsh Government. *Rights, Respect, Equality: Statutory Guidance for Governing Bodies of Maintained Schools*; Welsh Government: Cardiff, UK, 2019.
41. Russell, W.; Stenning, A. Beyond active travel: Children, play and community on streets during and after the coronavirus lockdown. *Cities Health* **2020**, *5*, S196–S199. [CrossRef]
42. Clemens, V.; Deschamps, P.; Fegert, J.M.; Anagnostopoulos, D.; Bailey, S.; Doyle, M.; Eliez, S.; Hansen, A.S.; Hebebrand, J.; Hillegers, M.; et al. Potential effects of "social" distancing measures and school lockdown on child and adolescent mental health. *Eur. Child Adolesc. Psychiatry* **2020**, *29*, 739–742. [CrossRef]