International Journal of Population Data Science

Journal Website: www.ijpds.org

Mother and Infant Research Electronic Data Analysis (MIREDA): A protocol for creating a common data model for federated analysis of UK birth cohorts and the life course

Mike Seaborne[1,2,3,*], Hope Jones[1,3], Neil Cockburn[4], Stevo Durbaba[5,6], Arturo González-Izquierdo[4], Amy Hough[7], Dan Mason[7], Carlos Sánchez-Soriano[8], Chris Orton[9], Armando Méndez-Villalon[10], Tom Giles[10], David Ford[9], Phillip Quinlan[10,11], Krish Nirantharakumar[4,12], Lucilla Poston[5,13,6], Rebecca Reynolds[8], Gillian Santorelli[7], and Sinead Brophy[1,14,15,3,16]

Submission History

Submitted:	09/02/2024
Accepted:	24/06/2024
Published:	12/09/2024

[1]National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Swansea, UK
[2]Data Lab, National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Swansea, UK
[3]Born in Wales, National Centre for Population Health and Wellbeing Research, Swansea University Medical School, Swansea, UK
[4]Institute of Applied Health Research, University of Birmingham, Birmingham, UK
[5]eLIXIR, Born in South London, King's College London, Strand, London, UK
[6]King's College London, School of Life Course and Population Sciences, London, UK
[7]Born in Bradford, Bradford Institute for Health Research, Bradford Royal Infirmary, Bradford, UK
[8]Born in Scotland, Centre for Cardiovascular Science, University of Edinburgh, Queen's Medical Research Institute, Edinburgh, UK
[9]Health Data Science, Swansea University, Swansea, UK
[10]Digital Research Service, University of Nottingham, Nottingham, UK
[11]School of Medicine, University of Nottingham, Nottingham, UK
[12]Mum-PreDiCT, University of Birmingham, Birmingham, UK
[13]Department of Women and Children's Health, School of Life Course and Population Sciences, King's College London, London, UK
[14]Health Data Research UK, London, UK
[15]Administrative Data Research Wales, Swansea University Medical School, Swansea, UK
[16]Health and Care Research Wales, Cardiff, UK

Abstract

Introduction
Birth cohorts are valuable resources for studying early life, the determinants of health, disease, and development. They are essential for studying life course. Dynamic longitudinal electronic cohorts use routinely collected data, are live, and can reduce selection bias specifically associated with direct recruitment in traditional birth cohorts. However, they are limited to health and administrative data and may lack contextual information.

The MIREDA (Mother and Infant Research Electronic Data Analysis) partnership creates a UK-wide birth cohort by aligning existing electronic birth cohorts to have the same structure, content, and vocabularies, enabling UK-wide federated analyses.

Objectives
1) Create a core dynamic, live UK-wide electronic birth cohort with approximately 500,000 new births per year using a common data model (CDM).

2) Provide data linkage and automation for long-term follow up of births from the Clinical Practice Research Datalink (CPRD), MuM-PreDiCT and the 'Born in' initiatives of Bradford, Wales, Scotland, and South London for comparable analyses.

Methods
We will establish core data content and collate linkable data. A suite of extraction, transformation, and load (ETL) tools will be used to transform data for each birth cohort into the CDM. Transformed datasets will remain within each cohort's trusted research environment (TRE). Metadata will be uploaded for the public to the Health Data Research (HDRUK) Innovation Gateway. We will develop a single online data access request for researchers. A cohort profile will be developed for researchers to reference the resource.

Ethics
Each cohort has approval from their TRE through compliance with their project application processes and information governance.

Dissemination
We will engage with researchers in the field to promote our resource through partnership networking, publication, research collaborations, conferences, social media, and marketing communications strategies.

Keywords
birth cohort; life course; data science; data linkage; population data; routinely collected health data; unified medical language system

*Corresponding Author:
Email Address: m.j.seaborne@swansea.ac.uk (Mike Seaborne)

https://doi.org/10.23889/ijpds.v9i2.2406

Introduction

Rapid socio-economic changes, including recent increases in cost of living in the United Kingdom (UK), contribute to widening inequalities that impact population health [1]. These changes, compounded by the strain on the National Health Service (NHS), social care services and education disruptions post-COVID-19 pandemic are expected to adversely affect maternal, infant and child health outcomes [2]. They have strong negative associations with deprivation and poor diet [3], unhealthy maternal body mass index (BMI) [4, 5], shorter intervals between pregnancies [6], pregnancy in adolescence [7], negative health behaviours [8], and conditions such as anaemia [9]. These alone can be responsible for deteriorating maternal, infant and child health and wellbeing, and can be damaging to development, future health and prosperity of offspring through their life course [10].

Effective monitoring and evaluation of interventions are essential to mitigate these risks and enhance health and wellbeing across affected populations.

Technology advancements and medical treatments continue to improve population health; however, a greater emphasis on informed preventative measures is needed to alleviate the burden on healthcare resources [11]. Optimising maternal, infant and child health during the perinatal period and beyond promises substantial benefits in the preventing illness and improvinged overall health.

Anonymised, routinely collected data and advanced data linkage techniques offer unprecedented insights into health trends over time, including pre- and post-pandemic shifts and their impacts on health outcomes, service provisions, and policy development.

Traditional birth cohorts have been pivotal in studying early life determinants of health, disease and development [12]. While they provide invaluable longitudinal data, direct recruitment methods limit representation and may not capture individuals born outside specific recruitment parameters [12–14].

Life course studies from birth are invariably complex as there are multiple, often interacting, factors which may influence outcomes. Therefore, the type and content of data collected is extensive and may include maternal characteristics, pregnancy complications - physical and mental health, birth outcomes, and other longitudinal data from social care, developmental milestones, education, household composition, etc.

We are living in an era with increasing access to vast amounts of health data and with growing analytical expertise, to interpret and disseminate meaningful insights about our nations' health. This provides an exciting new capacity for further life course research. During the pandemic large amounts of health data was brought together, but significant issues emerged when comparing routinely collected data across different UK nations [15]. Difficulties were compounded by a lack of standardisation of clinical coding systems, terminology, definitions, formats, difference in type and/or quantity of data, and data access governance [16] and, thus barriers to accessible, comparable data.

The evolving field of data science has facilitated the development of large-scale, anonymised data repositories within trusted research environments (TREs) [17, 18]. Data linkage techniques help to integrate diverse data sources for anonymised individuals to enhance research capabilities for comprehensive and representative population studies. They aid removal of obstacles such as the need to recruit, and serve as a means by which a representative population cohort can grow with the population, ensuring future relevance.

Despite existing national core datasets, which provide general statistical outputs, the granularity and versatility offered by anonymised data linkage techniques are unmatched. Such an anonymised and linkable single UK birth dataset is not currently available. and such centralised birth repository is not available for all four UK. Several electronic birth cohorts compile routinely collected data using diverse approaches to analyse population-level health across the life course. Each has similar data but with varying linkable data support [19–23]. These cohorts have emerged in tandem with improved accessibility, reduced research barriers, and the need for population-specific analyses that may not be generalised across dissimilar populations.

Harmonisation of these cohorts is a goal that aligns with the Findable, Accessible, Interoperable, Reusable (FAIR) Guiding Principles for scientific data management and stewardship [24]. The principles aim to establish comparability and consistency of analyses through a widely accepted standard. Achieving this would result in a network of cohorts gathering uniform core data, augmented with additional linkable datasets to extend the breadth and depth of research possibilities. Harmonisation will not only improve the findability and accessibility of birth data across our nations but will also enhance interoperability and reusability of the data for diverse research needs. This satisfies the FAIR criteria and will significantly advance the field of life course research.

The established standards of the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) will help us to achieve consistency. It standardises both structure and content (terminologies, vocabularies, and coding schemes) of observational data and has been implemented in other health data analyses to produce reliable, comparable analyses. The CDM contains 37 standardised tables relating to clinical data and vocabularies. There are additional domains for health systems, health economics, derived elements, and metadata (16).

The Mother and Infant Research Electronic Data Analysis (MIREDA) partnership aims to establish a unified platform for dynamic UK-wide birth cohorts using OMOP CDM. It will consolidate harmonised data into common core datasets for 4.35 million live births since January 1st, 2014. There will be ongoing updates and long-term follow-up and it will grow by approximately 500,000 births per year. Data will remain within TREs, ensuring privacy and security while facilitating federated analyses via the platform for improving the UK's infant and maternal health. TREs' will link MIREDA core data to public health, neonatal health, imaging, primary and secondary care, etc., and users will not be exposed to the raw data. 'Dummy' datasets will be created to evaluate researchers' analysis code before applying it to the real data.

Access will be facilitated by Health Data Research UK (HDR-UK) and governed by their TRE principles and best practices [25]. Access to MIREDA cohorts will be streamlined through a centralised HDR-UK online application. Metadata will be available to all via the Health Data Research (HDRUK)

Innovation Gateway. Initial birth cohorts include: Born in Wales (BiW) [20], Born in Scotland (BiS) [23], Born in Bradford (BiBBS/BiB4ALL) [26–28], early Life data Cross-Linkage in Research (eLIXIR) partnership (Born in South London) [21], and Clinical Practice Research Datalink (CPRD, for England) with linkage to Multimorbidity in Pregnancy: Determinant, Consequences, Clusters and Trajectories (MuM-PreDiCT) [22]. MuM-PreDiCT locations span four UK nations, but we will initially use only England's data.

This phase will run for one year and is scheduled for completion in July 2024.

Methods

Study design

The MIREDA partnership will assemble standardised, harmonised birth cohorts relating to mother, baby, and child as a federated common data framework. This will include structured definitions, descriptions and documentation of the data using the OMOP CDM, a comprehensive data dictionary, a data quality plan, and data governance policy. It will be available for creating a sustainable and reactive health system to inform public health decisions and policy.

Birth cohorts

The project unites the cohorts from CPRD (and MuM-PreDiCT), BiW, eLIXIR (BiSL), BiBBS/BiB4ALL, and BiS. Combined, they represent approximately 4.35 million live births. The CPRD alone consists of approximately four million births in England. Approximate proportional birth contributions for each are: 92%, 7%, 1%, 0.3%, and 0.1%, respectively. MuM-PreDiCT will represent a proportion of CPRD births with maternal multimorbidity (Table 1), estimated between 16-25% [29].

The proportional data contributions largely stem from England and Wales representing nationwide live births, with England's population nearly twenty times that of Wales and over five times the combined population of the other three. Scotland's cohort, limited to births from 2020, is not nationwide. The smaller cohorts from Bradford and South London include all local live births, and that offer rich ethnic diversity details missing in the other cohorts.

MuM-PreDiCT data for the remaining three nations will follow in subsequent work. We estimate that the combined cohorts will amass an additional 500,000 births per year.

Where available, live births on or after January 1st, 2014, will be identified from routine data.

Core data

MIREDA will collaborate to review the data accessible to each birth cohort and establish a consensus for their common data. This will be used to form a mutually agreed core dataset to be implemented for each cohort and will include the following dataset modules: maternal socio-demographics, baby socio-demographics, pregnancy, post-birth maternal health, birth, infancy and school age (Supplementary Table 1).

Linkable data

In addition to the core data, cohorts will have linkable datasets available to them which may include health visitor data, education, primary and secondary care, congenital abnormalities, vaccinations, breastfeeding, surveys, census, child protection and more (Supplementary Table 2).

Data storage

Each electronic birth cohort will retain person-level data, outcomes, and exposures within their respective trusted research environment (TRE), adhering strictly to its established rules.

Data tools

A suite of extraction, transformation and load (ETL) tools (WhiteRabbit [30], Convenient and Reusable Rapid OMOP Transformer (CaRROT) Mapper [31], and CaRROT-CDM [32] will be used to restructure concepts into the OMOP CDM.

Table 1: Birth cohort descriptions and data contribution

Cohort	Description	Start date	Initial inclusion	Est. prospective annual contribution
CPRD (MuM-PreDiCT)	All live births in England (Multimorbid mothers in CPRD)	01/01/2014	4,000,000 (640,000–1,000,000)	400,000 (64,000–100,000)
BiW	All live births in Wales	01/01/2014	290,000	29,000
eLIXIR	All live births from S. London	01/10/2018	50,000	14,000
BiBBS/BiB4ALL	All live births in Bradford	01/01/2014	15,000	6,000
BiS	Recruited live births: NHS Borders and NHS Lothian, Scotland All live births: NHS Lothian.	01/01/2020 10/06/2024	4,000	8,000

Data extraction and transformation

When data wrangling is complete and the core datasets assembled, each site will scan their data using WhiteRabbit (24,25) to produce a scan report. The anonymised reports will contain detailed information about tables, fields, and values that appear in fields. The reports will be extracted from the TREs for mapping to OMOP and creation of transformation rules.

Transformation rules will be imported into the TREs as .json files and will be used by local installations of the CaRROT-CDM tool to transform the core datasets into the OMOP CDM.

Data access

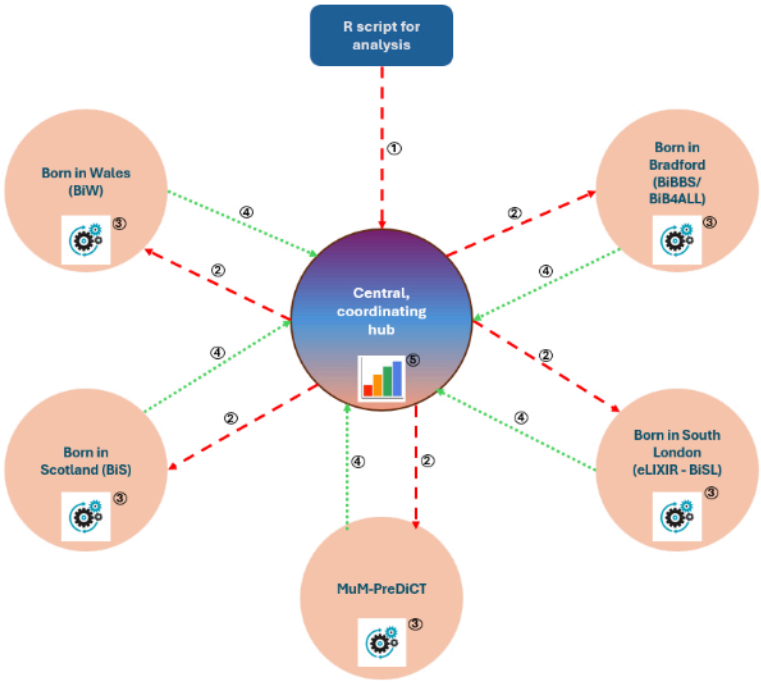
Data access will be granted by central request via the Health Data Research (HDRUK) Innovation Gateway. Here, the MIREDA collection will be available. It will detail the metadata for the cohorts and their linkable datasets. We will adapt the current HDRUK data access request (DAR) form to satisfy each TRE's requirements and create a single access request form for the collection, streamlining the application process to provide prompt access for rapid output analysis. No person-level data will leave any TRE. Instead, access will be used to create anonymised, aggregated tables and results of statistical analyses within each TRE. To do so, common R scripts will be developed and sent from a central hub to each TRE (step 1), where they will be implemented locally to produce the output required. They will be inspected by local analysts (step 2) before transferring to the central hub (step 3) for synthesis and further analysis (step 4), Figure 1.

Discussion

This protocol outlines the construction of a federated UK longitudinal birth cohort, leveraging routine data from UK healthcare providers, structured in a multi-modular format encompassing various stages from pregnancy to early childhood. The adoption of the OMOP CDM for this initiative is pivotal, ensuring the FAIR principles are embedded within the data management framework. By defining core variables and incorporating supplementary data from diverse sources, the protocol facilitates the enhancement and contextualisation of core datasets. Utilising the OMOP CDM provides standardisation of data, enabling uniform analyses across cohorts. It fosters the use of a single, centralised coding strategy across different TREs. Consequently, this approach permits the execution of a single analysis script across multiple cohorts, obviating the necessity for distinct scripts for each dataset. This will save time and costs, without compromising the quality of data and data linkage already available to the TREs.

The harmonised datasets created under this CDM will empower network studies for integrated research across different centres while ensuring that TREs maintain control over their data. This is crucial for upholding the security protocols and governance standards of TREs. Moreover, the central management of access requests via HDR-UK

Figure 1: Federated analysis infrastructure. The numbers in the figure describe the following steps



1. Analysis R script sent to central data hub (TRE) with specifications for the data model and analyses.
2. These are relayed to each birth cohort's TRE.
3. R scripts are run on each TRE's own system to create anonymised data tables and statistics.
4. These are sent back to the central hub for analysis.
5. Central hub compiles and analyses the data.

ensures a consistent application process that meets the requirements of each TRE. It aligns with the FAIR principles by providing a transparent, efficient, and universally applicable approach to data access, thereby advancing the integrity and utility of research conducted on these valuable cohorts.

Strengths and limitations of this study

Strengths

Establishes a comprehensive, UK-wide database which consolidates clinical data from maternity, neonatal, child health and education records. It will be enriched with quantitative and qualitative results from survey data to provide context and insight to the routinely collected data.

It will be a streamlined, central access point for the convenience of access request and rapid output.

The methodology uses an established standard for data harmonisation and standardisation through use of the OMOP CDM to create comparable core datasets for research.

Limitations

Routine, administrative data may be missing or prone to errors, it may also lack context. Relocation outside of cohorts' remits may lead to data loss and will be censored accordingly.

Proportional data representation seems skewed toward English data as England's population is approximately 20 times larger than Wales, and the other cohorts represent smaller, regional populations. Not all cohorts are able to contribute data as far back as 2014. Scotland's cohort is smaller as it is newer, with data only available from 2020, is restricted to two NHS regions at present, and initially only recruited through opt-in. Both London and Bradford are significantly smaller, However, they represent ethnically diverse populations, which is not available to the others.

Currently, we cannot include Northern Ireland's data, but future work will develop this too.

Future directions

OMOP CDM is an internationally accepted standard, so future development will include working with other cohorts, and other countries for international research collaboration. MIREDA will expand the OMOP standard to other non-healthcare data using a pseudo-OMOP style solution. Differences between cohorts will be leveraged to extrapolate findings where comparable data in others is missing. MIREDA will use pooled cohort data to tackle rare conditions, aiming to holistically prevent and mitigate risk factors impacting child health. Collaborating with HDR-UK, we will develop educational resources and establish a library of coding scripts for efficient data analysis.

Conclusion

OMOP CDM is internationally recognised as a healthcare data standard with robust training and support infrastructure, which is ideal for standardising data collaboratively and fostering enhanced research networks. It enables rapid, consistent data comparisons across regions, countries, and other localities. Supported by a variety of software tools and expertise from the Observational Health Data Sciences and Informatics (OHDSI) group, OMOP CDM is readily adoptable in TRE environments, ensuring continuity with existing policies, procedures and standards.

Acknowledgements

Supported by grants: MRC Partnership [MR/X02055X/1], MatCHNet pump-priming [U20005/302873] and MRC Programme [MR/X009742/1].

Statement of conflicts of interest

No known competing interests are known among any members of the MIREDA partnership involved in this study.

Ethics statement

Data access complies with the information governance requirements of each TRE. Anonymised data ensures compliance with the Data Protection Act 2018, providing only aggregated data and statistics. Each TRE holds ethical approval, obviating the need for additional ethical approval beyond standard project approval procedures.

Dissemination

The MIREDA partnership will actively engage, communicate, and disseminate this new data resource among researchers in maternal and infant health. We will promote resources, tools, collaborative expertise and infrastructure through publications, research collaborations, conferences, social media/marketing strategies.

MIREDA will present at relevant health and health data research organisation meetings and events. Workshops at targeted conferences will promote the linked data to encourage grant applications and research collaborations.

Face-to-face workshops will showcase research using MIREDA data, with recordings uploaded as online webinars. Key group invitees include policymakers, NHS healthcare professionals in maternity, and industry stakeholders.

We will develop MOOCs (massive open online courses) to enhance skills in maternal and infant linked data analysis. Regular blogs and podcasts will discuss data utilisation.

We will collaborate with international partners to raise awareness of MIREDA for research collaborations.

MIREDA information will be hosted on the websites of MIREDA partners, and resources will be promoted through ADR Digital Insights.

Training will be provided in association with HDRUK training group, ADR training and NIHR to develop capacity and expertise. Pump priming funding will enable early career researchers to use the data.

Patient and public involvement

Patients and/or the public participated in the design, or conduct, or reporting, or dissemination plans of this research. We follow the Co-production of Research and Strategy (CORDS) standard operating procedure and the UK standards for PPI involvement and National Institute for Health Research (NIHR) guidance from INVOLVE [33]. Records of PPI activity are maintained using the Public Involvement in Research Impact Toolkit (PIRIT) [34]. Each cohort and co-connect works with its own PPI groups under different names/guises to contribute to each step of this process.

Data availability statement

Data will be available upon reasonable request through the Health Data Research (HDRUK) Innovation Gateway.

Full URLs contained within manuscript

Health Data Research Innovation Gateway https://www.healthdatagateway.org/.
Observational Health Data Sciences and Informatics https://www.ohdsi.org/.

References

1. Broadbent P, Thomson R, Kopasker D, Mccartney G, Meier P, Richiardi M, et al. The public health implications of the cost-of-living crisis: outlining mechanisms and modelling consequences. Lancet Reg. Health - Eur. 2023;27:100585. https://doi.org/10.1016/j.lanepe.2023.100632

2. Fears R, Gillett W, Haines A, Norton M, Meulen V ter. Post-pandemic recovery: use of scientific advice to achieve social equity, planetary health, and economic benefits. Lancet Planet. Health 2020;4:e383–4. https://doi.org/10.1016/S2542-5196(20)30176-5

3. Haggarty P, Campbell DM, Duthie S, Andrews K, Hoad G, Piyathilake C, et al. Diet and deprivation in pregnancy. Br. J. Nutr. 2009 Nov.;102:1487–97. https://doi.org/10.1017/S0007114509990444

4. Black RE, Allen LH, Bhutta ZA, Caulfield LE, Onis M de, Ezzati M, et al. Maternal and child undernutrition: global and regional exposures and health consequences. Lancet Lond. Engl. 2008;371:243–60. https://doi.org/10.1016/S0140-6736(07)61690-0

5. Nurul-Farehah S, Rohana AJ. Maternal obesity and its determinants: A neglected issue? Malays. Fam. Physician Off. J. Acad. Fam. Physicians Malays. 2020;15:34.

6. Schummers L, Hutcheon JA, Hernandez-Diaz S, Williams PL, Hacker MR, Vanderweele TJ, et al. Association of Short Interpregnancy Interval With Pregnancy Outcomes According to Maternal Age. JAMA Intern. Med. 2018 Dec.;178:1661–70. https://doi.org/10.1001/JAMAINTERNMED.2018.4696

7. Chen XK, Wen SW, Fleming N, Demissie K, Rhoads GG, Walker M. Teenage pregnancy and adverse birth outcomes: a large population based retrospective cohort study. Int. J. Epidemiol. 2007 Apr.;36:368–73. https://doi.org/10.1093/IJE/DYL284

8. Smoking and pregnancy patient information leaflet | RCOG [Internet]. Available from: https://www.rcog.org.uk/for-the-public/browse-our-patient-information/smoking-and-pregnancy-patient-information-leaflet/.

9. Nair M, Churchill D, Robinson S, Nelson-Piercy C, Stanworth SJ, Knight M. Association between maternal haemoglobin and stillbirth: a cohort study among a multi-ethnic population in England. Br. J. Haematol. 2017 Dec.;179:829–37. https://doi.org/10.1111/BJH.14961

10. Mikkelsen B, Williams J, Rakovac I, Wickramasinghe K, Hennis A, Shin HR, et al. Life course approach to prevention and control of non-communicable diseases. BMJ [Internet] 2019 Jan.;364. Available from: https://www.bmj.com/content/364/bmj.l257 https://www.bmj.com/content/364/bmj.l257.abstract. https://doi.org/10.1136/BMJ.L257

11. Hochlaf D, Quilter-Pinner H, Kibasi T. The case for a new approach to public health and prevention: The progressive policy think tank. 2019; Available from: www.ippr.org

12. O'connor M, Spry E, Patton G, Moreno-Betancur M, Arnup S, Downes M, et al. Better together: Advancing life course research through multi-cohort analytic approaches. Adv. Life Course Res. [Internet] 2022;53. Available from: http://creativecommons.org/licenses/by/4.0/. https://doi.org/10.1016/j.alcr.2022.100499

13. Downs JM, Ford T, Stewart R, Epstein S, Shetty H, Little R, et al. An approach to linking education, social care and electronic health records for children and young people in South London: a linkage study of child and adolescent mental health service data. Available from: http://bmjopen.bmj.com/. https://doi.org/10.1136/bmjopen-2018-024355

14. Overy C, Reynolds LA, Tansey EM, Group H of BR. History of the Avon longitudinal study of parents and children (ALSPAC), c. 1980–2000: the transcript of a Witness Seminar held by the History of Modern Biomedicine Research Group, Queen Mary, University of London, on 24 May 2011. 2012;122.

15. COVID-19 Data and Connectivity [Internet]. HDR UK [cited 2024 26]; Available from: https://www.hdruk.ac.uk/covid-19-data-and-connectivity/.

16. Data Standardization – OHDSI [Internet]. Available from: https://www.ohdsi.org/data-standardization/.

17. Lugg-Widger FV, Angel L, Cannings-John R, Hood K, Hughes K, Moody G, et al. Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: Managing the morass. Int. J. Popul. Data Sci. 3:432. https://doi.org/10.23889/ijpds.v3i3.432

18. Price G, Peek N, Eleftheriou I, Spencer K, Paley L, Hogenboom J, et al. An overview of Real-World Data infrastructure for cancer research. Clin. Oncol. [Internet] 2024 [cited 2024 26]; Available from: https://www.sciencedirect.com/science/article/pii/S0936655524001080.

19. Wright J, Small N, Raynor P, Tuffnell D, Bhopal R, Cameron N, et al. Cohort Profile: The Born in Bradford multi-ethnic family cohort study. Int. J. Epidemiol. 2013;42:978–91. https://doi.org/10.1093/ije/dys112

20. Jones H, Seaborne MJ, Kennedy NL, James M, Dredge S, Bandyopadhyay A, et al. Cohort profile: Born in Wales—a birth cohort with maternity, parental and child data linkage for life course research in Wales, UK. BMJ Open 2024 Jan.;14:e076711. https://doi.org/10.1136/BMJOPEN-2023-076711

21. Carson LE, Azmi B, Jewell A, Taylor CL, Flynn A, Gill C, et al. Cohort profile: the eLIXIR Partnership-a maternity-child data linkage for life course research in South London, UK. BMJ Open 2020;10:39583. https://doi.org/10.1136/bmjopen-2020-039583

22. Lee SI, Eastwood KA, Moss N, Azcoaga-Lorenzo A, Subramanian A, Anand A, et al. Protocol for the development of a core outcome set for studies of pregnant women with pre-existing multimorbidity. BMJ Open 2021 Oct.;11:e044919. https://doi.org/10.1136/BMJOPEN-2020-044919

23. Born in Scotland | The University of Edinburgh [Internet]. Available from: https://www.ed.ac.uk/cardiovascular-science/born-in-scotland

24. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 2016 Mar.;3:160018. https://doi.org/10.1038/sdata.2016.18

25. Alliance UHDR, NHSX. Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems. 2021 Dec.; Available from: https://zenodo.org/record/5767586. https://doi.org/10.5281/ZENODO.5767586

26. Dickerson J, Bridges S, Willan K, Kelly B, Moss RH, Lister J, et al. Born in Bradford's Better Start (BiBBS) interventional birth cohort study: Interim cohort profile. Wellcome Open Res. 2022 Oct.;7:244. https://doi.org/10.12688/wellcomeopenres.18394.1

27. BiB4ALL [Internet]. Born Bradf. [cited 2024 13]; Available from: https://borninbradford.nhs.uk/what-we-do/cohort-studies/born-in-bradford/

28. Cohort studies: BiB4ALL [Internet]. 2024 Feb.; Available from: https://borninbradford.nhs.uk/what-we-do/cohort-studies/born-in-bradford/

29. Lee SI, Azcoaga-Lorenzo A, Agrawal U, Kennedy JI, Fagbamigbe AF, Hope H, et al. Epidemiology of pre-existing multimorbidity in pregnant women in the UK in 2018: a population-based cross-sectional study. BMC Pregnancy Childbirth 2022 Dec.;22:1–15. https://doi.org/10.1186/S12884-022-04442-3/TABLES/5

30. WhiteRabbit for ETL design – OHDSI [Internet]. Available from: https://www.ohdsi.org/analytic-tools/whiterabbit-for-etl-design/

31. About - CaRROT-Mapper [Internet]. Available from: https://hdruk.github.io/CaRROT-Docs/CaRROT-Mapper/about/

32. About - CaRROT-CDM [Internet]. Available from: https://hdruk.github.io/CaRROT-Docs/CaRROT-CDM/About/

33. INVOLVE. Guidance on co-producing a research project [Internet]. 2018 Feb.;1–20. Available from: https://www.invo.org.uk/posttypepublication/guidance-on-co-producing-a-research-project/

34. Public Involvement in Research Impact Toolkit (PIRIT) - Marie Curie Research Centre - Cardiff University [Internet]. Available from: https://www.cardiff.ac.uk/marie-curie-research-centre/patient-and-public-involvement/public-involvement-in-research-impact-toolkit-pirit15

Abbreviations

MIREDA:	Mother and Infant Research Electronic Data Analysis
OMOP:	Observational Medical Outcomes Partnership
CDM:	Common Data Model
TRE:	Trusted Research Environment
BiW:	Born in Wales
eLIXIR:	early Life data Cross-Linkage in Research
BiSL:	Born in South London
BiB4All:	Local name for BaBi Bradford
BaBi Bradford:	Born and Bred in Bradford
BiBBS:	Born in Bradford Birth Study
BiS:	Born in Scotland
MuM-PreDiCT:	Multimorbidity in Pregnancy: Determinant, Consequences, Clusters and Trajectories
ETL:	Extraction, Transformation, and Load
HDR-UK:	Health Data Research UK
GP:	General Practitioner
CPRD:	Clinical Practice Research Datalink
HES:	Hospital Episode Statistics
NIV:	National Immunisation Vaccine
DAR:	Data Access Request
MOOC:	Massive Open Online Courses
WHO:	World Health Organization
NIHR:	National Institute for Health Research
OHDSI:	Observational Health Data Sciences and Informatics
PPI:	Patient and Public Involvement
FAIR:	Findable, Accessible, Interoperable, Reusable

Supplementary Materials

Supplementary Table 1: Core data for birth cohort dataset modules

Dataset modules	Data	Additional information
Maternal socio-demographics	Week/Date of birth	
	Ethnicity	
	Deprivation Quintile	By locality at pregnancy start
	Rural/urban	By locality at pregnancy start
	Education	
	Qualifications	
	Occupation	
	Date of death	
Baby socio-demographics	Week/Date of birth	
	Sex	
	Ethnicity	
	Deprivation quintile	By locality at birth and school age
	Rural/urban residence	By locality at birth and school age
	Disability codes	
	Date of death	
Pregnancy	Date of antenatal visit	
	Maternal age	At booking visit/conception
	Maternal smoking status	
	Maternal Height	
	Maternal Weight	
	Maternal BMI	
	Existing diagnosis codes	Comorbidities existing at booking visit
	New event type	New events since previous visit e.g. diagnosis, procedure, medicines,
	New event code	etc.
	New event date	
	New event source	Event sources e.g. GP/hospital.
	Data of expected Delivery	
	Folic acid supplements indicator	
	Domestic violence indicator	
	WHOOLEY score	Or other depression indicators.
	Gravidity	
	Parity	
	Previous C-Section indicator	
	Date of previous C-section	
	Previous preterm birth indicator	
	Date of previous preterm birth	
	Gestational age	At visit.
Post-birth (maternal)	Postnatal depression indicator	
	Event type	Events within 10 weeks of birth; may include: diagnosis, procedure,
	Event code	medication, etc.
	Event date	
	Event source	Event sources e.g. GP/hospital.
Birth	Date of delivery	
	Gestational age	At birth.
	Birth weight	
	Event type	Events from birth to 28 days; may include: diagnosis, procedure,
	Event code	medication, etc.
	Event date	
	Event source	Event sources e.g. GP/hospital.
	Mode of delivery	E.g. c-section, forceps, vaginal, etc.

Continued

Supplementary Table 1: Continued

Dataset modules	Data	Additional information
Infancy	Event type	As above but from 29 days to school age
	Event code	
	Event date	
	Event source	
	Neonatal admission indicator	
	Neonatal event indicator	
	Breastfeeding indicator	Breastfeeding, mixed/formula indicator
	Breastfeeding duration	
School age	Event type	As above but for events from school start age to 12 years.
	Event code	
	Event date	
	Event source	

Table 2: Linkable datasets

	Born in Wales	eLIXIR: (BiSL)	Born in Scotland	Born in Bradford		CPRD
				BiBBS	BiB4All	
Core data						
Pregnancy, birth, infancy.	Yes	Yes	Yes	Yes	Yes	Yes
Routine data						
Health Visitor	Yes	Will have	Will have	Yes	Yes	-
Education	Yes	Yes	Children too young but will have in the future		Children too young at present	-
GP	Yes	Yes	Will have	Yes	Yes	Yes
Hospital admission	Yes	Yes	Will have	Yes	Yes	Yes
A&E	Yes	Will have (HES)	Will have	Yes	Yes	-
Congenital abnormalities	Yes	Yes	Will have	Yes	Yes	
National community child health (vaccination, breastfeeding)	Yes	Health visitor data for breastfeeding. NIV data for vaccination - applied for	Will have	Yes	Yes	
Survey with family						
Pregnancy	Yes	No		Yes	Yes	-
18 months	Yes	No				
3-5 years	Yes	No				
Primary school	Yes	No				
Biological samples						
Bloods	-	Yes	Yes	Yes	Yes	
Other datasets						
Census 2011	Yes	Could apply for this				-
Census 2021	Yes	As above				-
Looked after children	Yes	As above				
Child protection	Yes	As above				
Child in receipt of care and support	Yes	As above				