International Journal of Population Data Science

Journal Website: www.ijpds.org

IJPDS International Journal of Population Data Science

Swansea University Prifysgol Abertawe

Open science and phenotyping in UK administrative health, education and social care data: the ECHILD phenotype code list repository

Matthew A. Jay[1,*], Kate Lewis[1], Difei Shi[1], Rebecca Langella[1], Tony Stone[1,2], Sorcha Ní Chobhthaigh[3], Ania Zylbersztejn[1], Ruth Blackburn[1], and Katie Harron[1]

Submission History

Submitted:	11/11/2024
Accepted:	01/04/2025
Published:	13/05/2025

[1]University College London Great Ormond Street Institute of Child Health, 30 Guilford Street, London, WC1N 1EH
[2]Sheffield Centre for Health and Related Research, The University of Sheffield, Sheffield, UK
[3]University College London Institute of Global Health, 30 Guilford Street, London, WC1N 1EH

Abstract

Administrative health data, such as the Hospital Episode Statistics (HES), can be used to identify groups of people with a particular target condition, a process known as phenotyping. Clinical phenotypes are useful as exposures, covariates and outcomes in research studies using administrative data, including health data linked to other sources such as the Education and Child Health Insights from Linked Data (ECHILD) project. ECHILD brings together HES and other national health datasets with the National Pupil Database and children's social care data for all of England as a data asset that can be accessed by researchers at UK institutions. Because using linked administrative data is complex, the ECHILD team has created additional resources to improve the accessibility of ECHILD. One such initiative is the ECHILD Phenotype Code List Repository. The Repository is a fully open and searchable website containing phenotype code lists that can be used in ECHILD and beyond. As well as a primer on phenotyping, it includes summaries of each code list and R and Stata implementation scripts. The Repository was designed according to a set of principles to ensure that finding and using code lists is easy and standardised. The ECHILD Phenotype Code List Repository is a step forward in the findability and use of phenotype code lists in ECHILD and its constituent datasets.

Keywords
ECHILD; open science, phenotyping; repository

*Corresponding Author:
Email Address: matthew.jay@ucl.ac.uk (Matthew A. Jay)

https://doi.org/10.23889/ijpds.v10i2.2943

Introduction

Administrative health data usually contain diagnostic and procedure information recorded according to national standards. For example, Hospital Episode Statistics (HES) admitted patient care, collated by National Health Service (NHS) England, records all NHS-funded acute inpatient hospital activity in England [1, 2]. Coders record diagnostic information using the 10th edition of the International Statistical Classification of Diseases and Related Health Problems (ICD-10), and procedure information using the Office for Population Censuses and Surveys Classification of Interventions and Procedures, version 4 (OPCS-4) [3]. Analysts, in turn, can use this information to identify groups of patients recorded with a given condition or undergoing certain procedures. For instance, users interested in severe congenital heart defects might use ICD-10 and OPCS-4 codes to identify children in HES with this group of conditions [4]. This process of health status identification is referred to as phenotyping. The targeted feature (e.g., severe congenital heart defects) is called the phenotype [5] and the list of codes used to identify that phenotype is the phenotype code list.

While subject to certain limitations, use of such information in whole-population datasets enables research that would not be possible with primary data collection methods. This is largely because of the near-whole-population coverage of administrative data. Linkage to other datasets, such as the National Pupil Database (NPD) [6], which includes data on all children in state schools in England, can prove even more powerful in enabling cross-sectoral analyses. This has been achieved by the Education and Child Health Insights from Linked Data (ECHILD) project [7, 8]. ECHILD, which also includes a range of other health and social care datasets for all of England, is currently the only national-level linkage between these datasets and includes information on individuals born since September 1984 [7, 8].

Phenotyping in administrative data is essential as it is often the only way in such studies to identify health exposures and outcomes. The process of defining a phenotype and creating a code list, however, takes time [9]. Ideally, this should involve consultation with clinicians and other stakeholders such as coders and patient groups, as well as empirical validation against a gold standard (e.g., positive predictive value and negative predictive value, among other metrics [5]) and through checking temporal and geographical variation in coding in the target data. This is not always feasible or at least consumes time that would otherwise be spent on the research problem. Fortunately, many teams have developed and published code lists, though they can be difficult to find and implement as they are usually published as non-machine-readable tables in supplementary material. This introduces a risk of error as researchers must implement the list either by creating their own machine-readable file or by hard-coding the code list into their scripts.

To address this, we launched the ECHILD Phenotype Code List Repository [10], a website containing phenotype code lists that can be used in ECHILD and beyond (https://code.echild.ac.uk). In the spirit of promoting the FAIR Principles of Findability, Accessibility, Interoperability, and Reuse of data [11], the ECHILD team has been working to develop a community of practice supported by "How To" guides (https://howto.echild.ac.uk) and documentation, and now the Repository, which also includes example R and Stata scripts. In this article, we introduce the Repository, its rationale and its potential uses.

The ECHILD Phenotype Code List Repository is part of an ecosystem of repositories, such as CALIBER [12], the Health Data Research (HDR) UK Phenotype Library [13], the London School of Hygiene and Tropical Medicine's Data Compass [14] and OpenSAFELY's OpenCodelists [15], many of which have grown organically by teams working on a variety of datasets. We in turn developed the ECHILD repository to make it easier to find code lists relevant to ECHILD's health, education and social care data, and, crucially, to implement them using the example scripts. Repository entries are curated by the ECHILD team, meaning they are directly relevant to ECHILD. Early experience and informal feedback from users have shown that the availability of an easily searchable database of relevant code lists and implementation code has significantly reduced the difficulty of finding and using phenotype code lists in ECHILD.

Overview and examples

The Repository website contains a primer on phenotyping, details of our design principles, information on how to use the example scripts and the code lists and implementation scripts. The primer provides an overview of diagnosis and procedure coding in HES (as these are the most used fields in such code lists) as well as some caveats to using them, particularly those from other jurisdictions. Each code list is detailed on its own page.

Currently the Repository contains 11 code lists [4, 16–25] (Table 1). Most are generic covering several conditions (e.g., chronic health conditions of Hardelid et al. [16]), or more specific covering a particular group of conditions (e.g., the stress-related presentations of Blackburn et al. [20] and Ní Chobhthaigh et al. [21]). One currently covers a single condition: asthma (Lut et al. [22]). We also included a code list to identify emergency admissions as a number of code lists are designed only for use with such admissions [25].

It is possible to search the website and constituent lists by keyword or code. The latter is useful, for example, where researchers wish to know whether and in what context a known code has been used by others. For example, J45 encodes asthma. If searched on the website, it can be seen that J45 appears in Hardelid et al's [16] list of chronic health conditions without qualification and in Lut et al's [22] list for asthma but with 12 codes for other respiratory conditions that invalidate the asthma codes.

To provide detail of one example, consider the chronic health conditions list by Hardelid et al. [16], which can be accessed at https://code.echild.ac.uk/chc_hardelid_v1. Users will first see Repository details such as the code list name, version number, ID and links to download the code list file and associated R script and Stata do file. Below this, details of the code list are available, including its authors and citation (Hardelid et al. [16]), its target phenotype (chronic health conditions) and original purpose (to describe chronic health conditions among children who die), information on its creation and validation, groups and subgroups (nine groups

Table 1: Overview of phenotype code lists currently in the ECHILD phenotype code list repository

Authors (location)	**Target phenotype**	**Code system**	**Original purpose**	**Original format**	**Alterations in repository**
Hardelid et al. (England) [16]	Chronic health conditions	ICD-10	Describe what chronic health conditions children die with, using HES inpatient records as well as death certificates.	Table in supplementary material (PDF, not machine-readable).	None.
Cohen et al. (Canada) [17]	Medical complexity	ICD-10	Ascertain prevalence of medical complexity.	Table in supplementary material (PDF, not machine-readable).	Ten ICD-10-CA codes truncated to first 4 characters and 20 ICD-9 codes removed.
Feudtner et al. (USA) [18]	Complex chronic conditions	ICD-10	An update to an ICD-9 code list designed to study patterns of paediatric mortality and end-of-life care.	Table in supplementary material (Word, not machine-readable) & hard-coded in a Stata do file and SAS script.	This code list makes extensive use of ICD-10-CM and ICD-10-PCS codes, which are not used in HES. The repository version has therefore had 1,102 codes removed and 195 truncated.
Fraser et al. (England) [19]	Life-limiting and life-threatening conditions	ICD-10	Estimate the prevalence of life-limiting and life-threatening conditions in children.	Table in supplementary material (PDF, not machine-readable).	None.
Blackburn et al. (England) [20]	Stress-related presentations	ICD-10	Study stress-related presentations to hospital among adolescents.	Table in supplementary material (Word, not machine-readable).	None.
Ní Chobhthaigh et al. (England) [21]	Stress-related presentations	ICD-10	Update Blackburn et al's list and study stress-related presentations to hospital among adolescents.	Table in supplementary material (Word, not machine-readable).	None.
Herbert et al. (England) [23]	Adversity-related injuries	ICD-10	Study adversity-related admissions in adolescence. The accidental injuries group served as a comparison group.	Table in supplementary material (PDF, not machine-readable).	Code I42.1 (obstructive hypertrophic cardiomyopathy) was included in the original documentation but this should be I42.6 (alcoholic cardiomyopathy). In the repository, I42.1 is removed and replaced with I42.6.
Gimeno et al. (England) [4]	Severe congenital heart defects	ICD-10 OPCS-4	Describe trends in 5-year mortality among children with severe congenital heart defects compared to other children.	Table in supplementary material (PDF, not machine-readable).	In Gimeno et al's paper, OPCS-4 codes L02.1 to L02.9 and L03.1 are duplicated in eTable 2, first without a flag indicating required birthweight and gestational age, and second with this flag. Gimeno has confirmed (personal correspondence) that they should appear once WITH this flag.

Continued

Table 1: Continued

Authors (location)	Target phenotype	Code system	Original purpose	Original format	Alterations in repository
Ford et al. (England) [24]	Anorectal malformations	ICD-10 OPCS-4	Estimate the birth prevalence of isolated and complex anorectal malformations, maternal risk factors and 5-year mortality.	Table in supplementary material (PDF, not machine-readable).	Two procedure codes were removed as these were incorrectly labelled as OPCS-4 codes in the original paper when they are in fact codes in the Clinical Coding & Schedule Development Group's list [26] used by private providers.
Lut et al. (International) [22]	Asthma	ICD-10	Cross-country comparison of asthma rates.	Written in prose in main text of paper (PDF, not machine-readable).	None.
NHS England (England) [25]	Emergency admissions	Admission method	Identify emergency admissions.	HES data dictionary (Excel, not machine-readable).	None.

HES Hospital Episode Statistics; ICD-10 the International Statistical Classification of Diseases and Related Health Problems, 10th edition; ICD-10-CA ICD-10 Canada; ICD-10-PCS ICD-10 Procedure Coding System; NHS National Health Service; OPCS-4 the Office of Population Censuses and Surveys Classification of Interventions and Procedures, 4th version.

by body system plus subgroups), alterations made to the list (none) and the number of codes ($n = 1{,}372$). A section on flags highlights special conditions that must be applied. In this case, some codes are only valid where admission length is at least three days and some others where the child is aged at least ten years. Finally, a preview of the code list is available.

The formatting of code lists

We designed the Repository according to a set of principles, detailed on the website, ensuring that all code lists are formatted equally and machine-readable. Code lists, all CSV files, at a minimum have five columns that provide the code, the dataset to which it relates, the field in that dataset, the code type (ICD-10 or OPCS-4) and the code description. Some lists have additional columns to indicate special conditions.

Each CSV file contains one code per row. Ranges (e.g., "D80-D89") are not permitted as these are not easily machine-readable. Additionally, this may cause problems where a code that would logically fall within the range does not currently exist but this is not apparent from the code list itself. An example is in Feudtner et al. [18], which includes the range D80 to D89 in the original paper, but D85, D87 and D88 do not exist. All code lists are arranged in the same order as the original publication, and use the same group labels, to aid cross-checking.

Only final code lists are admitted to the Repository with an associated publication where the list was first published. This is to ensure that the Repository contains original lists that have undergone at least some form of peer review, also ensuring consistency across studies using code lists.

Alterations to code lists

There are three types of necessary revision: to ensure the code list is compatible with ECHILD, to rectify errors in the original list or to make necessary updates to a list over time.

In the first case, alterations are necessary because only the standard ICD-10 is currently used in HES. The Clinical Modification (ICD-10-CM) [27] and Canadian version (ICD-10-CA) [28] (and others, though so far, no code list in the Repository uses any) contain 5-character codes. These provide more detailed classifications than the standard ICD-10. Where these are used (e.g., in Cohen et al's [17] and Feudtner et al's [18] lists of complex conditions), we truncate to the corresponding 4-character version. It is also sometimes necessary to remove Procedure Coding System (ICD-10-PCS) and ICD-9 codes.

The second change (fixing errors) has occurred rarely. For example, in Herbert et al's [23] list of adversity-related admissions, code I42.1 (obstructive hypertrophic cardiomyopathy) was erroneously included in the original but this should have been I42.6 (alcoholic cardiomyopathy). In the Repository, code I42.1 is removed and replaced with I42.6. Similarly, some codes appear twice in the original supplementary material to Gimeno et al's [4] paper on severe congenital heart defects. In the first instance, these are included erroneously without a restriction flag and in the second, correctly with the flag. The Repository reflects the correct version. Alterations (removing or truncating codes and fixing errors) is done by ECHILD staff prior to uploading to the Repository and all are documented on the website and in the code list files. Users must decide whether altered code lists remain valid for their purposes.

The third type of alteration refers to wholesale updates. Compare, for example, the code list of stress-related presentations by Blackburn et al. [20] with that of Ní Chobhthaigh et al. [21]. The latter is an update of the former, accounting for the latest research and practice in paediatric mental health. Both lists are available separately in the Repository. Users may wish to refer to a previous version because it aligns more closely with their research objectives or to evaluate published research. Where the ECHILD team is aware of an updated version that is not yet included in the Repository, we will include a note on the old version notifying users.

Submitting a code list

We will add relevant code lists on an on-going basis and we welcome submissions from the research community, who should consult the website to find out how. Code lists must be applicable to ECHILD's health, education or social care data, conform to our design principles and be a published, final version. Once formatted, code lists are checked by two ECHILD data scientists or researchers who independently and manually review the list to ensure it includes, and only includes, all codes correctly. This also involves checking group labels and flags as well as any alterations as described above. To date, all included code lists are applicable to HES inpatient records. We hope that, as users begin to explore other ECHILD datasets, such as the mental health, NPD and social care data, that more lists applicable to these datasets will be developed and admitted to the Repository.

Strengths and considerations

We believe that the ECHILD Phenotype Code List Repository is a step forward in the findability and use of phenotype code lists in ECHILD. The Repository could be used to support projects in a range of disciplines, including those not using ECHILD.

Our design principles and checks ensure that code lists are accurate, easy to find and implementable in the data. The example R and Stata scripts provide worked examples of implementation in real ECHILD data using a minimal number of packages (data.table [29] and RODBC [30]) in the case of R or only base functionality in the case of Stata. The Repository and scripts also aid understanding and dealing with special conditions on particular codes (e.g., codes that are only valid where the child is at least a certain age).

There are, however, some considerations to be borne in mind. The ECHILD team cannot determine whether a code list is appropriate for any given study. Researchers must assess this on a case-by-case basis, which involves considering the codes contained (and not contained) in a list and its possible sensitivity and specificity, which cannot be measured directly in ECHILD, vis-à-vis their target phenotype. This should include consideration of whether and how the code list was originally validated [5, 9] and how clinicians and coders record diagnostic information. This may vary over time and in different hospitals. While the Repository website provides summaries, users should consult original publications for details. Researchers are also recommended to carry out checks for coding depth (e.g., the number of codes recorded per hospital episode) and temporal and geographical variation in the frequency of codes. This is to assess possible changes in coding practice not due to changes in underlying incidence or prevalence of the target phenotype.

In future, the 11th ICD revision, ICD-11, which differs significantly from ICD-10, will become mandatory in the NHS, though timescales are currently unclear [31]. It will become necessary to develop and validate new code lists. The new coding system and code lists may affect estimates of prevalence and incidence over time due to changes in coding rather than underlying epidemiology.

Creating a sustainable future for ECHILD and other national data assets requires continued support for infrastructure. Establishing the Repository required a data scientist (who was also a researcher using ECHILD in applied projects) dedicated to the task. Going forward, staff time will be required when adding new code lists due to the Repository's design principles and quality checks. In other words, on-going maintenance of linked data assets, such as ECHILD, and their associated support materials, such as the Repository, require funding in addition to that earmarked for core research tasks.

Conclusion

The ECHILD Phenotype Code List Repository was established to make it easier to find, understand and implement phenotype code lists in ECHILD. It is freely available for the entire research community to use. We hope that it will also serve as a template for future researchers developing code lists in terms of how to format and release them, with an eye to machine-readability.

Acknowledgements

ECHILD is supported by ADR UK (Administrative Data Research UK), an Economic and Social Research Council (part of UK Research and Innovation) programme (ES/V000977/1, ES/X000427/1 and ES/X003663/1). AZ was supported by NIHR Programme Development Grant (NIHR206946). We would like to thank the ECHILD programme team and community. The ECHILD Database uses data from the Department for Education (DfE) and data provided by patients and collected by the National Health Service as part of their care and support. We gratefully acknowledge all children and families whose de-identified data are used in ECHILD. The DfE does not accept responsibility for any inferences or conclusions derived by the authors. We are grateful to the Office for National Statistics for providing the trusted research environment for the ECHILD database. The use of the Office for National Statistics statistical data does not imply the endorsement of the Office for National Statistics in relation to the interpretation or analysis of the statistical data.

Statement of conflicts of interest

None declared.

Ethics statement

Ethical approval was not required. This manuscript does not report a scientific study. The ECHILD Phenotype Code List Repository does not need ethical approval as it is not a study and only contains non-confidential code lists and associated material.

Data availability statement

This manuscript does not report a scientific study and has no data to report. The ECHILD Phenotype Code List Repository is available at https://code.echild.ac.uk.

References

1. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). International Journal of Epidemiology. 2017;46(4):1093-i. https://doi.org/10.1093/ije/dyx015

2. NHS England. Hospital Episode Statistics (HES). 2024. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics (accessed 24 September 2024).

3. NHS England. The NHS Classifications Browser. 2023. Available from: https://classbrowser.nhs.uk/#/ (accessed 26 June 2024).

4. Gimeno L, Brown K, Harron K, Peppa M, Gilbert R, Blackburn R. Trends in survival of children with severe congenital heart defects by gestational age at birth: A population-based study using administrative hospital data for England. Paediatric and Perinatal Epidemiology. 2023;37(5):390-400. https://doi.org/10.1111/ppe.12959

5. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. Journal of Clinical Epidemiology. 2011;64(8):821-9. https://doi.org/10.1016/j.jclinepi.2010.10.006

6. Jay MA, Mc Grath-Lone L, Gilbert R. Data Resource: the National Pupil Database (NPD). International Journal of Population Data Science. 2019;4(1). https://doi.org/10.23889/ijpds.v4i1.1101

7. ECHILD. ECHILD. 2024. Available from: https://www.echild.ac.uk/ (accessed 24 September 2024).

8. Mc Grath-Lone L, Libuy N, Harron K, Jay MA, Wijlaars L, Etoori D, et al. Data Resource Profile: The Education and Child Health Insights from Linked Data (ECHILD) Database. International Journal of Epidemiology. 2021;51(1):17-f. https://doi.org/10.1093/ije/dyab149

9. Matthewman J, Andresen K, Suffel A, Lin L, Schultze A, Tazare J, et al. Checklist and guidance on creating codelists for routinely collected health data research. NIHR Open Research. 2024;4(20). https://doi.org/10.3310/nihropenres.13550.2

10. ECHILD. ECHILD Phenotype Code List Repository. 2024. Available from: https://code.echild.ac.uk/ (accessed 2 September 2024).

11. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3(1):160018. https://doi.org/10.1038/sdata.2016.18

12. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. Journal of the American Medical Informatics Association. 2019;26(12):1545–59. https://doi.org/10.1093/jamia/ocz105

13. HDRUK. HDRUK Phenotype Library. 2024. Available from: https://phenotypes.healthdatagateway.org/ (accessed 21 October 2024).

14. London School of Hygiene and Tropical Medicine. LSHTM Data Compass. 2024. Available from: https://datacompass.lshtm.ac.uk/ (accessed 21 October 2024).

15. OpenSAFELY. OpenCodelists. 2024. Available from: https://www.opencodelists.org/ (accessed 21 October 2024).

16. Hardelid P, Dattani N, Gilbert R. Estimating the prevalence of chronic conditions in children who die in England, Scotland and Wales: a data linkage cohort study. BMJ Open. 2014;4(8):e005331. https://doi.org/10.1136/bmjopen-2014-005331

17. Cohen E, Berry JG, Camacho X, Anderson G, Wodchis W, Guttmann A. Patterns and costs of health care use of children with medical complexity. Pediatrics. 2012;130(6):e1463-70. https://doi.org/10.1542/peds.2012-0175

18. Feudtner C, Feinstein JA, Zhong W, Hall M, Dai D. Pediatric complex chronic conditions classification system version 2: updated for ICD-10 and complex medical technology dependence and transplantation. BMC Pediatr. 2014;14:199. https://doi.org/10.1186/1471-2431-14-199

19. Fraser LK, Miller M, Hain R, Norman P, Aldridge J, McKinney PA, Parslow RC. Rising National Prevalence of Life-Limiting Conditions in Children in England. Pediatrics. 2012;129(4):e923-e9. https://doi.org/10.1542/peds.2011-2846

20. Blackburn R, Ajetunmobi O, Mc Grath-Lone L, Hardelid P, Shafran R, Gilbert R, Wijlaars L. Hospital admissions

for stress-related presentations among school-aged adolescents during term time versus holidays in England: weekly time series and retrospective cross-sectional analysis. BJPsych Open. 2021;7(6):e215. https://doi.org/10.1192/bjo.2021.1058

21. Ní Chobhthaigh S, Jay M, Blackburn R. Emergency hospital admissions for stress-related presentations among secondary school-aged minoritised young people in England. Br J Psych. 2025;226(2):63-71. https://doi.org/10.1192/bjp.2024.123

22. Lut I, Lewis K, Wijlaars L, Gilbert R, Fitzpatrick T, Lu H, et al. Challenges of using asthma admission rates as a measure of primary care quality in children: An international comparison. Journal of Health Services Research & Policy. 2021;26(4):251-62. https://doi.org/10.1177/13558196211012732

23. Herbert A, Gilbert R, González-Izquierdo A, Li L. Violence, self-harm and drug or alcohol misuse in adolescents admitted to hospitals in England for injury: a retrospective cohort study. BMJ Open. 2015;5(2):e006079. https://doi.org/10.1136/bmjopen-2014-006079

24. Ford K, Peppa M, Zylbersztejn A, Curry JI, Gilbert R. Birth prevalence of anorectal malformations in England and 5-year survival: a national birth cohort study. Archives of Disease in Childhood. 2022;107(8):758. https://doi.org/10.1136/archdischild-2021-323474

25. NHS England. Hospital Episode Statistics Data Dictionary. 2024. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary (accessed 24 September 2024).

26. The Clinical Coding & Schedule Development Group. The Clinical Coding & Schedule Development Group. n.d. Available from: https://www.ccsd.org.uk/ (accessed 21 October 2024).

27. National Center for Health Statistics. ICD-10-CM Files. 2024. Available from: https://www.cdc.gov/nchs/icd/icd-10-cm/files.html (accessed 24 September 2024).

28. Canadian Institute for Health Information. Canadian Coding Standards for ICD-10-CA and CCI. 2024. Available from: https://secure.cihi.ca/estore/productSeries.htm?pc=PCC189 (accessed 24 September 2024).

29. Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T, Schwendinger B. data.table: Extension of 'data.frame' [R package]. 2024. Available from: https://cran.r-project.org/web/packages/data.table/index.html (accessed 1 November 2024).

30. Ripley B, Lapsley M. RODBC: ODBC Database Access. 2023. Available from: https://cran.r-project.org/web/packages/RODBC/index.html (accessed 1 November 2024).

31. NHS England. Terminology and Classifications. 2024. Available from: https://digital.nhs.uk/services/terminology-and-classifications (accessed 23 October 2024).

Abbreviations

CSV:	Comma-separated values
ECHILD:	Education and Child Health Insights from Linked Data
HES:	Hospital Episode Statistics
ICD-9/ICD-10/ICD-11:	International Statistical Classification of Diseases and Related Health Problems, 9th/10th/11th edition
ICD-10-CA:	ICD-10, Canada.
ICD-10-CM:	ICD-10, Clinical Modification.
ICD-10-PCS:	ICD-10 Procedure Coding System.
NHS:	National Health Service
NPD:	National Pupil Database
OPCS-4:	Office of Population Censuses and Surveys Classification of Interventions and Procedures, 4th revision