PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	COHORT PROFILE : THE NATIONAL CONGENITAL ANOMALY REGISTRATION DATASET IN ENGLAND
AUTHORS	Broughan, Jennifer; Wreyford, Ben; Martin, Danielle; Melis, Gabriella; Randall, Kay; Obaro, Ewoma; Broggio, John; Aldridge, Nicholas; Stoianova, Sylvia; Johnson, Chloe; Gibbard, Donna; Stevens, Sarah; Fleming, Kate M

VERSION 1 – REVIEW

REVIEWER	Battersby, Cheryl Imperial College London, Neonatal Medicine
REVIEW RETURNED	29-Aug-2023

GENERAL COMMENTS	A very useful and informative paper that I enjoyed reviewing. Thank you. Overall it is a very useful description of a cohort profile and believe it warrants publication. I have a few queries to confirm my own understanding, and suggestions to improve the manuscript and look forward to the authors' responses. This is a descriptive paper on a cohort of babies with congenital anomalies (CAs) held on the National Congenital Anomaly and Rare Disease Registration Service (NCARDRS). These infants were born and cared for in England. The authors consist of the team that manage the database at the National Disease Registration Service (NDRS) NHS England in Data and Analytics Directorate. The manuscript documents the changes in population coverage over time. NCARDRS coverage increased from 22% of total births in England in 2015 to 100% national coverage in 2018. Prior to 2015, data collection was performed independently by regional registers in England rather than nationally. NCARDRS registers CAs that occur in babies born alive and stillborn, fetal losses and terminations in England. Data sources include records from secondary and tertiary health care providers with maternal or paediatric departments, private providers and laboratories covering fetal medicine, maternity or paediatric services. The authors suggest that the cohort is the largest globally with approximately 600,000 births per year. 21,000 new registrations of babies or fetuses with suspected or confirmed CAs added each year.

The manuscript details the data sources, data structure, data processing, findings to date (headlines only), and in the discussion makes references to some of the uses of the data to date.

Future plans of the cohort include linkage to other datasets.

Suggestions/queries
Nicely written to provide an overview of the important findings.
“With a coverage of approximately 600,000 total births per year, the NCARDRS CA register is the largest globally.”

This can be misinterpreted
For clarity, please also add to this the 21,000 number of anomalies on the register so its clear.

Strengths and limitations section

“NCARDRS is one of the largest congenital anomaly registers in the world. Since 2018, coverage of congenital anomalies has been national across England (approximately 600,000 total births per year). This enables the calculation of accurate estimates of prevalence even for rare congenital anomalies. Legacy congenital anomaly registration data is held for some regions for births since 1985.

Suggest for clarity that these are births versus the number of babies with congenital anomalies each year to add “21,000 ” figure

INTRODUCTION

Informative introduction outlining the significant contribution of congenital anomalies to death globally. In England and Wales, congenital anomalies is the most common cause of death in the post neonatal period (36 % of deaths).

Question: can the authors please clarify if the cohort/dataset also registered with Health Data Research (HDR) UK cohort registry to enhance visibility?

https://www.healthdatagateway.org/about/cohort-discovery

COHORT description

Outlines inclusion and exclusion criteria. Useful supplementary table 1 supplementary table 1 summarising the inclusion criteria, exclusions, differences with EUROCAT, enhanced registration.

Figure 1: Useful diagram of maps to illustrate population coverage increasing over time and 100% after 2018

- Please can the authors include the colour key for brown and make the legend bigger as it is very small and difficult to read.

Registration model and source data:

Figure 2. Very useful schematic describing the multisource registration process used for congenital anomalies.

However, there are some phrases that are less well known to experts in the field and I would suggest to explain “waterfall”. Data waterfall in the manuscript says this is a semi-automatic process that is used to input data into the data management sytem. Can the authors please provide more detail on this and also annotate the figure to explain this.

“CARA- the anomaly is confirmed according to set criteria”- Can authors please explain what this is too. I cannot find the abbreviation spelt out in full in the rest of the manuscript. Who confirms the anomaly according to “set criteria” ? What is this set criteria? This is important as it is a useful validation step to ensure confidence in the validity and accuracy of data considering the first step includes suspected anomalies. Can the authors please provide more details about this?

Suggestion for additions to Figure 2: Is it also possible please to include the actual organisations that contribute data as these are familiar e.g. HES, ONS? And include what identifiers are used for linkage e.g. NHS number or date of birth, full name and address. I think these are useful to add in the diagram for an overall summary.

Data processing:

Please can the authors add more specific clarification regarding whether names and identifiers are removed from the free text before registration officers review these data? Do you use Natural language processing to interrogate free text data or is this manually done by the registration officers?

I have suggested this clarification as parent and patient groups are interested the handling of personal identifiable information.

Regulatory details
Ethical and governance arrangements:
I note the legal regulatory framework that has granted the use of the NCARDRS.
Can the authors please provide more detail about the steps that external applicants who wish to use the data need to take to appy for data, including the requirement to apply for REC for individual studies?

I would suggest to expand the data availability statement within the manuscript to include a detailed section on useful information for external researchers or clinicians accessing/ applying for the data and the steps they need to take.

Data structure
Can the authors please include example of the 5 main tables (with dummy data) in the supplementary so the readers can follow what is described in this section ?

Key data fields: Suggest when referring to the 15 conditions, “enhanced registration” as a term is used again as I believe this is the reason these 15 conditions are monitored?

I see congenital diaphragmatic hernia (CDH) included in the 15 conditions – in supplementary S1 this is under congenital malformations of other parts of the musculoskeletal system. However, the impact of CDH is largely respiratory (the diaphragm) and therefore should be under “respiratory system anomalies” – and under respiratory system anomalies , it says “no enhanced registration”.

Can the authors elaborate on the reason why CDH is not under respiratory anomalies and how this affects reporting on respiratory anomalies i.e likely under reporting if categories are assigned as it is?

Findings to date :
As of June 2023, NCARDRS held information on 117,682 mothers and 121,184 babies born in England since 2015.

Table 1 – please can the authors elaborate more on Table 1 which is the first table presenting results over time and number of anomalies registered on NCARDRS:
1) “Total number of confirmed and probable anomalies” – can the confirmed anomalies be reported separately to the probable please as there is , I believe a validation process to confirm diagnoses? I presume the confirmed will be less than the probable?

2) “Birth population with active congenital anomaly registration” - can you please add a denote specifically for this row to explain these numbers further as these are many times more the yearly numbers? Is this the prevalence as it’s cumulative? If so, does it imply that many die over the years hence reducing numbers?

Table 2:
I would suggest to move the “test” column to another separate table rather than be included in Table 2. Table 2 is meant to be a summary of key data items available for each congenital anomaly registration. However, the denotes 1 for “test” information states that test information (test date, type, results, provider etc) are only consistently registered for conditions with enhanced registration. These only include 15 conditions. Please confirm this is correct.

Typo line 38 page 12 , extra full stop within the brackets

Risk factor information: Ethnicity is mentioned and available. Is consanguinity also available? Deprivation score also available. Can the authors please include detail of any work that has been conducted examining the completeness of ethnicity or consanguinity or maternal/paternal age, and also geography, and also the impact of assisted conception (as this is also available) please as these risk factors are relevant to congenital anomalies.

Strengths
I think the combination of multi sources, lab and clinical data is a key strength and should be included in strengths.

	Limitations – these are intertwined with the strengths. For the structure, can the authors please separate strengths from limitations? Future work: Suggest to include other potential new data sources mentioned that may be added in addition to maternity services dataset in NHS England? Any steps to include primary care data? Neonatal unit data example Natioanl Neonatal Research Database linkage? Would the authors be welcome to more collaborations or applications from external researchers for data? Perhaps this could be highlighted in the discussion. Also suggest authors can discuss the implications of genetic advances and increasing use of whole genome sequencing, and how this may affect NCARDRS. Conclusions: I would suggest whilst there are many strengths of NCARDRS, providing data along the life course, it should be noted that NCARDRS contain information related to health only. Suggest to highlight that future work should include linkage to social and education data.

REVIEWER	Zylbersztejn, Ania University College London Great Ormond Street Institute of Child Health , Population, policy and practice
REVIEW RETURNED	09-Sep-2023

GENERAL COMMENTS	Thank you for inviting me to review the data resource profile for the National Congenital Anomaly and Rare Disease Registration Service (NCARDRS) database. NCARDRS collects data on rare diseases and congenital anomalies (CAs) in live births, stillbirths, fetal loses and terminations of pregnancy in England. CAs can be notified from multiple sources (and across child's life course) to ensure high case ascertainment. Since 2018, NCARDRS has national coverage. The rates of anomalies by subtype match those reported by EUROCAT for Europe. This is extremely valuable resource for epidemiological and public health studies and this data profile will be a helpful reference for researchers interested in using NCARDRS for research. The profile covers information on how data are collected and processed, a comprehensive overview of findings to data and a good overview of strengths and weaknesses. I have no major comments on structure or content, but I have a few suggestions for clarifications for authors to consider: 1) Some additional explanations/clarification would be useful for non-UK audience or audiences who eg.: do not know about administrative data in England. I am not suggesting that authors write anything detailed, just add a brief context. Some of the things I noted are: - Explaining that England has universal healthcare - You mention UK rare disease strategy (Introduction, paragraph 3), a brief explanation/context of what this is could be helpful - You mention maternity services dataset (as future work), could you provide a reference and maybe a brief overview of what it is / what types of information is covered?

	- Similarly HES data (mentioned in registration model and data source) – perhaps you can explain that it covers all hospital admission records for England and add a reference 2) More detail on key information covered by NCARDRS would be useful for researchers interested in using this dataset. E.g.: Table 2 could be revised to have one row per variable, with separate columns for variable name, definition (where relevant) & which dataset it is included in (& maybe coverage or years available if relevant). Or instead, if data dictionary is available somewhere, you could provide a reference? 3) Related to above: is linkage to HES and ONS available for all records? I.e. babies and mothers? Or just records where a relevant congenital anomaly is recorded to increase case ascertainment? if for all, maybe you could mention the types of information available with this linkage (eg causes, timing and place of death; details of hospital admissions including diagnoses and procedures, outpatient records?) 4) What denominator population data is available and where it comes from? Do you receive individual-level data from ONS birth registration? or aggregated data on number of births per region / sex (any other characteristics)? Can you provide a reference to birth registration from ONS (e.g. for numbers in table 1) This is a minor point, but in the abstract you say “with a coverage of approximately 600,000 total births per year” – this gives the impression that you collect data on all births in England (rather than anomalies in all births) so perhaps more accurate to rephrase as “With surveillance covering approximately 600,000 total births….” 5) Regional registers prior to 2015: In introduction you mention that registration became established from 60s and 70s in many countries. Is this when regional registers were established in England too? In the “strengths and limitations” bullet points you mention that legacy data goes back to 1985, so it might be worth explaining somewhere in the manuscript the coverage of legacy data also maintained by NCARDRS (if it’s also available for research via NCARDRS) 6) Can you provide additional information on linkage – what is the linkage process for linking to other datasets? Is it deterministic based on steps? Who links the data? 7) You mention probable and confirmed CAs – what happens to the probable cases? Are they kept in the data? or removed if not confirmed? is that indicator available in the dataset?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. Cheryl Battersby, Imperial College London
Comments to the Author:
A very useful and informative paper that I enjoyed reviewing. Thank you.

Overall it is a very useful description of a cohort profile and believe it warrants publication. I have a few queries to confirm my own understanding, and suggestions to improve the manuscript and look forward to the authors' responses.

This is a descriptive paper on a cohort of babies with congenital anomalies (CAs) held on the National Congenital Anomaly and Rare Disease Registration Service (NCARDRS). These infants were born and cared for in England. The authors consist of the team that manage the database at the National Disease Registration Service (NDRS) NHS England in Data and Analytics Directorate.

The manuscript documents the changes in population coverage over time. NCARDRS coverage increased from 22% of total births in England in 2015 to 100% national coverage in 2018. Prior to 2015, data collection was performed independently by regional registers in England rather than nationally.

NCARDRS registers CAs that occur in babies born alive and stillborn, fetal losses and terminations in England. Data sources include records from secondary and tertiary health care providers with maternal or paediatric departments, private providers and laboratories covering fetal medicine, maternity or paediatric services.

The authors suggest that the cohort is the largest globally with approximately 600,000 births per year. 21,000 new registrations of babies or fetuses with suspected or confirmed CAs added each year.

The manuscript details the data sources, data structure, data processing, findings to date (headlines only), and in the discussion makes references to some of the uses of the data to date.

Future plans of the cohort include linkage to other datasets.

Suggestions/queries
Nicely written to provide an overview of the important findings.
"With a coverage of approximately 600,000 total births per year, the NCARDRS CA register is the largest globally."

This can be misinterpreted
For clarity, please also add to this the 21,000 number of anomalies on the register so its clear.

1. Agree this has potential for misinterpretation and we have done this and improved the clarity of this point.

Strengths and limitations section

"NCARDRS is one of the largest congenital anomaly registers in the world. Since 2018, coverage of congenital anomalies has been national across England (approximately 600,000 total births per year). This enables the calculation of accurate estimates of prevalence even for rare congenital anomalies. Legacy congenital anomaly registration data is held for some regions for births since 1985.

Suggest for clarity that these are births versus the number of babies with congenital anomalies each year to add "21,000 " figure

2. Agree this has potential for misinterpretation and we have done this and improved the clarity of this point.

INTRODUCTION

Informative introduction outlining the significant contribution of congenital anomalies to death globally. In England and Wales, congenital anomalies is the most common cause of death in the post neonatal period (36 % of deaths).

Question: can the authors please clarify if the cohort/dataset also registered with Health Data Research (HDR) UK cohort registry to enhance visibility?

https://www.healthdatagateway.org/about/cohort-discovery

3. NCARDRS is not currently registered with the Health Data Research (HDR) UK cohort registry but we will explore this

COHORT description

Outlines inclusion and exclusion criteria. Useful supplementary table 1 supplementary table 1 summarising the inclusion criteria, exclusions, differences with EUROCAT, enhanced registration.

Figure 1: Useful diagram of maps to illustrate population coverage increasing over time and 100% after 2018

- Please can the authors include the colour key for brown and make the legend bigger as it is very small and difficult to read.

4. A new Fig 1 has been uploaded with a more comprehensive legend, the addition to the colour key. We also took the opportunity to enhance the labelling and clarity of the images.

Registration model and source data:

Figure 2. Very useful schematic describing the multisource registration process used for congenital anomalies.

However, there are some phrases that are less well known to experts in the field and I would suggest to explain “waterfall”. Data waterfall in the manuscript says this is a semi-automatic process that is used to input data into the data management system. Can the authors please provide more detail on this and also annotate the figure to explain this.

5. We have added a description of the data water fall in the Data Processing section (pages 9-10). A new figure 2 has been uploaded with further explanation of the processes in plain English.

“CARA- the anomaly is confirmed according to set criteria”- Can authors please explain what this is too. I cannot find the abbreviation spelt out in full in the rest of the manuscript. Who confirms the anomaly according to “set criteria” ? What is this set criteria? This is important as it is a useful validation step to ensure confidence in the validity and accuracy of data considering the first step includes suspected anomalies. Can the authors please provide more details about this?

6. A new paragraph describing disease coding and confirmation of an anomaly has been added to the Data processing section (page 10) describing the anomaly status and the criteria used to define this status. We have removed references to CARA which is the name of our data management system but of only internal relevance.

Suggestion for additions to Figure 2: Is it also possible please to include the actual organisations that contribute data as these are familiar e.g. HES, ONS? And include what identifiers are used for linkage e.g. NHS number or date of birth, full name and address. I think these are useful to add in the diagram for an overall summary.

7. This has been done, more detail has been added to Figure 2 as mentioned above.

Data processing:

Please can the authors add more specific clarification regarding whether names and identifiers are removed from the free text before registration officers review these data? Do you use Natural language processing to interrogate free text data or is this manually done by the registration officers?

I have suggested this clarification as parent and patient groups are interested the handling of personal identifiable information.

8. This has been added and is a valuable clarification for the reasons stated. NLP is not currently used to interrogate free text data, these data are manually reviewed by a registration officer

Regulatory details
Ethical and governance arrangements:
I note the legal regulatory framework that has granted the use of the NCARDRS.
Can the authors please provide more detail about the steps that external applicants who wish to use the data need to take to appy for data, including the requirement to apply for REC for individual studies?
9. Access to NCARDRS congenital anomaly is either through DARS or data can be made available by working in partnership with NDRS. More detail has been added to the data availability section but data availability and governance required will depend on the circumstances of each applicant and type of study.

I would suggest to expand the data availability statement within the manuscript to include a detailed section on useful information for external researchers or clinicians accessing/ applying for the data and the steps they need to take.
10. We have added some additional data but access to detail depends on individual circumstances and direction to DARS is the first step.

Data structure
Can the authors please include example of the 5 main tables (with dummy data) in the supplementary so the readers can follow what is described in this section ?

11. This was an oversimplification of the data on our part, and so we have made changes to the text to avoid further misunderstanding. In reality data are held in 600+ tables which are transformed into approximately 100+ tables describing the data, but they can be grouped into those main themes

Key data fields: Suggest when referring to the 15 conditions, “enhanced registration” as a term is used again as I believe this is the reason these 15 conditions are monitored?
12. Agree, this has been done

I see congenital diaphragmatic hernia (CDH) included in the 15 conditions – in supplementary S1 this is under congenital malformations of other parts of the musculoskeletal system. However, the impact of CDH is largely respiratory (the diaphragm) and therefore should be under “respiratory system anomalies” – and under respiratory system anomalies , it says “no enhanced registration”.

Can the authors elaborate on the reason why CDH is not under respiratory anomalies and how this affects reporting on respiratory anomalies i.e likely under reporting if categories are assigned as it is?

13. We collect 1000+ congenital anomalies and conditions and we organised Table S1 roughly according to ICD-10/BPA system1 with some amendments to align with EUROCAT subgroup coding2, but allowing greater granularity. We have added a foot note to explain this.
EUROCAT guidance states CHD should be reported as a GI anomaly and so we have added it there and added another subgroup describing abdominal wall conditions to ensure international consistency and consistency with our own routine reporting. However, we have deviated from the EUROCAT subgroups and included some further granularity as well as the inclusion of some conditions that are collected by NCARDRS and not EUROCAT.
CDH will not be unreported as it is reported separately in prevalence tables for our routine reporting3
1World Health Organization. ICD-10: International Statistical Classification of Diseases and Related Health Problems. Geneva: World Health Organization; 2010.
2https://eu-rd-platform.jrc.ec.europa.eu/system/files/public/eurocat/Guide_1.5_Chapter_3.3.pdf
3 https://digital.nhs.uk/data-and-information/publications/statistical/ncardrs-congenital-anomaly-statistics-annual-data/

Findings to date :
As of June 2023, NCARDRS held information on 117,682 mothers and 121,184 babies born in England since 2015.

Table 1 – please can the authors elaborate more on Table 1 which is the first table presenting results over time and number of anomalies registered on NCARDRS:
1) “Total number of confirmed and probable anomalies” – can the confirmed anomalies be reported separately to the probable please as there is , I believe a validation process to confirm diagnoses? I presume the confirmed will be less than the probable?

14. Anomaly status and how they are determined is now described in the text in response to the comment above). Confirmed and probable anomalies have similar interpretation and are reportable in almost all situations. We do not report on suspected anomalies until enough evidence has been provided to promote their status. Probable is used for situations where the anomaly was highly likely but further evidence was not possible eg termination

2) “Birth population with active congenital anomaly registration” -can you please add a denote specifically for this row to explain these numbers further as these are many times more the yearly numbers? Is this the prevalence as it’s cumulative? If so, does it imply that many die over the years hence reducing numbers?

15. This row label has been changed to “Number of live and still births in regions with active congenital anomaly registration (denominator)”. This is the denominator we use to calculate birth

prevalence and reflects the increasing proportion of the birth population that was covered by congenital anomaly registration as NCARDRS opened new regions, eventually reaching national coverage (of all approx. 600K births in England).

Table 2:
I would suggest to move the "test" column to another separate table rather than be included in Table 2. Table 2 is meant to be a summary of key data items available for each congenital anomaly registration. However, the
denotes 1 for "test" information states that test information (test date, type, results, provider etc) are only consistently registered for conditions with enhanced registration. These only include 15 conditions. Please confirm this is correct.

16. You are correct that test information is only available for the 15 types of conditions audited by the FASP but this accounts for 35% of the data in 2020 (4501 out of 13065 babies had a FASP condition) so it is not a small proportion of the whole dataset. For this reason we have left Table 2 as it is but we have made changes in the text, clarifying that these reflect groups of conditions and that it is a substantial proportion of the data.

Typo line 38 page 12 , extra full stop within the brackets
17. Thank you, this has been corrected

Risk factor information: Ethnicity is mentioned and available. Is consanguinity also available? Deprivation score also available. Can the authors please include detail of any work that has been conducted examining the completeness of ethnicity or consanguinity or maternal/paternal age, and also geography, and also the impact of assisted conception (as this is also available) please as these risk factors are relevant to congenital anomalies.

18. Ethnicity, consanguinity, maternal age, assisted conception and deprivation score are all available and listed in Table 2. Location (post code of delivery/booking) is very well completed and this is how deprivation score is obtained this is also extremely well completed. There has to date been limited work on assessing the completion of other fields, and this is something we hope to address soon.

Strengths
I think the combination of multi sources, lab and clinical data is a key strength and should be included in strengths.

19. Agree, this has been added

Limitations – these are intertwined with the strengths. For the structure, can the authors please separate strengths from limitations?
20. This has been doneWe tried this but it was very disjointed. The strengths and limitations are intertwined and for this reason we have left the section as it was.

Future work:
Suggest to include other potential new data sources mentioned that may be added in addition to maternity services dataset in NHS England? Any steps to include primary care data? Neonatal unit data example Natioanl Neonatal Research Database linkage?

21. The Future Works section describes planned linkage to MSDS and refers to the potential of primary care data although this has yet to be realised. Linkage with the NNRD has been discussed, but NCARDRS already has access to neonatal clinical systems and extracts for most trusts in England. We are in discussions with the leaders of clinical audit datasets and NHS E highly specialised services and I have added these.

Would the authors be welcome to more collaborations or applications from external researchers for data? Perhaps this could be highlighted in the discussion.

22. This has been added. NCARDRS is subject to NHS E policy and process on data release but data is available via partnership working and we welcome applications or requests for collaboration as capacity allows.

Also suggest authors can discuss the implications of genetic advances and increasing use of whole genome sequencing, and how this may affect NCARDRS.

23. A sentence has been added to the new “Multisource” section in Strengths

Conclusions:
I would suggest whilst there are many strengths of NCARDRS, providing data along the life course, it should be noted that NCARDRS contain information related to health only. Suggest to highlight that future work should include linkage to social and education data.

24. Fair comment, this has been clarified and added.

Reviewer: 2
Dr. Ania Zylbersztejn, University College London Great Ormond Street Institute of Child Health
Comments to the Author:
Thank you for inviting me to review the data resource profile for the National Congenital Anomaly and Rare Disease Registration Service (NCARDRS) database. NCARDRS collects data on rare diseases and congenital anomalies (CAs) in live births, stillbirths, fetal loses and terminations of pregnancy in England. CAs can be notified from multiple sources (and across child’s life course) to ensure high case ascertainment. Since 2018, NCARDRS has national coverage. The rates of anomalies by subtype match those reported by EUROCAT for Europe. This is extremely valuable resource for epidemiological and public health studies and this data profile will be a helpful reference for researchers interested in using NCARDRS for research.

The profile covers information on how data are collected and processed, a comprehensive overview of findings to data and a good overview of strengths and weaknesses. I have no major comments on structure or content, but I have a few suggestions for clarifications for authors to consider:

1) Some additional explanations/clarification would be useful for non-UK audience or audiences who eg.: do not know about administrative data in England. I am not suggesting that authors write anything detailed, just add a brief context. Some of the things I noted are:
- Explaining that England has universal healthcare
Agree, and added
- You mention UK rare disease strategy (Introduction, paragraph 3), a brief explanation/context of what this is could be helpful
Agree, and added
- You mention maternity services dataset (as future work), could you provide a reference and maybe a brief overview of what it is / what types of information is covered?
Agree, and added
- Similarly HES data (mentioned in registration model and data source) – perhaps you can explain that it covers all hospital admission records for England and add a reference
Agree, and added

2) More detail on key information covered by NCARDRS would be useful for researchers interested in using this dataset. E.g.: Table 2 could be revised to have one row per variable, with separate columns for variable name, definition (where relevant) & which dataset it is included in (& maybe coverage or years available if relevant). Or instead, if data dictionary is available somewhere, you could provide a reference?
A table with date items listed by row would be 48 rows long and we feel that this is the best way to summarize the data. A detailed data dictionary can be made available on request to DARS and this is stated in the Data Availability Statement.
All the variables in Table 2 are in the NCARDRS congenital anomaly dataset and available for all years. We have made changes to the title of Table 2 to reflect this.

3) Related to above: is linkage to HES and ONS available for all records? I.e. babies and mothers? Or just records where a relevant congenital anomaly is recorded to increase case ascertainment? if for all, maybe you could mention the types of information available with this linkage (eg causes, timing and place of death; details of hospital admissions including diagnoses and procedures, outpatient records?)
As stated in the text we obtain a data extract from ONS for ascertaining those with a congenital anomaly listed on the death certificate. Linkage of all patients in the dataset is not currently available but is planned in the near future.

4) What denominator population data is available and where it comes from? Do you receive individual-level data from ONS birth registration? or aggregated data on number of births per region / sex (any other characteristics)? Can you provide a reference to birth registration from ONS (e.g. for numbers in table 1)
We have access to individual-level data from ONS birth registration and we aggregate this ourselves by IMD, Sex and region/area so we cannot provide a reference, this is analysis done in-house. This has been added to the Study Population section (page 8).

This is a minor point, but in the abstract you say "with a coverage of approximately 600,000 total births per year" – this gives the impression that you collect data on all births in England (rather than anomalies in all births) so perhaps more accurate to rephrase as "With surveillance covering approximately 600,000 total births…."
Agree, and have changed as above in response to Reviewer 1

5) Regional registers prior to 2015: In introduction you mention that registration became established from 60s and 70s in many countries. Is this when regional registers were established in England too?

In the "strengths and limitations" bullet points you mention that legacy data goes back to 1985, so it might be worth explaining somewhere in the manuscript the coverage of legacy data also maintained by NCARDRS (if it's also available for research via NCARDRS)

The introduction states that national registration was in place in 2018, there was a brief attempt at a national congenital anomaly register by ONS but this closed in 2010 – this has been added to the Introduction. Prior to this data collection was region. The scope of the cohort profile covers NCARDRS data collected from 2015 and we have not included the previous congenital anomaly registries for clarity. We do hold the data from the regional registries and I have added in a line in the Future Work section. It is not currently available via the DARS access route.

6) Can you provide additional information on linkage – what is the linkage process for linking to other datasets? Is it deterministic based on steps? Who links the data?
Linkage at the point of registration is deterministic based on NHS number, postcode, name and date of birth. This demographic identifiers are checked using NHS Spine. Further linkage for analytical purposes is a potential utility within the dataset – some datasets have been linked (eg HES and the NHSBA community prescriptions data) and linkage with MSDS is currently in development. Linkage is based on NHS number, date of birth and location information.

7) You mention probable and confirmed CAs – what happens to the probable cases? Are they kept in the data? or removed if not confirmed? is that indicator available in the dataset?
As above in response to reviewer 1, there is now a paragraph describing the anomaly status, which clarifies this point. Anomaly status describes the status of the anomaly and is listed in Table 2.

VERSION 2 – REVIEW

REVIEWER	Zylbersztejn, Ania University College London Great Ormond Street Institute of Child Health , Population, policy and practice
REVIEW RETURNED	27-Nov-2023

GENERAL COMMENTS	Authors addressed all of my comments and I have no further suggestions.