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.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

(This paper received three reviews from its previous journal but only two reviewers agreed to published their review.)

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cohort Profile: Children in Need Census (CIN) records of children referred for social care support in England
AUTHORS	Emmott, Emily H; Jay, Matthew A; Woodman, Jenny

VERSION 1 – REVIEW

REVIEWER	Anna Stewart School of Criminology and Criminal Justice Griffith University Australia
REVIEW RETURNED	20-Jun-2018

GENERAL COMMENTS	Cohort Profile: Children in Need Census (CIN) records of vulnerable children in England. This paper contains a description of the Children in Need (CIN) Census. It took me quite sometime to realise that this was a description of a data repository rather than a research study. The authors need to be clearer about the purpose of the paper and ensure that the reader is also clear about this purpose. I am assuming this paper is an attempt to provide metadata for the CIN and encourage researchers to consider this as a data source for future research. It is apparent from the description of the data there are a number of major drawbacks/issues in using these data. I think these should be clearly stated and discussed. These are the findings of this research not that there are 2.76 million children on the CIN as suggested in the abstract – this is a description of the data. The description of the children on the database is difficult to follow for someone from outside the UK. I have no idea what the baseline population is so I do not understand how many children this represents. I have no idea what proportion of children are in publicly funded schools. The Figure indicates that there are four databases (three overlapping circles and one subsumed) but I think there are only three CIN, Children in the School census and Children in Care. However, the names of these three databases keep changing through the document – absolute numbers would help. From what I can work out the issues with the data include

	1. Lack of a unique numeric identifier for a child. I understand from the text why this happens but I am not sure what this means for anyone who wishes to use the data. Does this mean there is no name or birthdate on the episodes? a. If so how can these data be linked to other data sets outside the ones described? For example how was it linked to the hospital episode statistics? The suggestion that 64% of children do not have one of the identifiers seems to make the data useless for examining educational outcomes. The authors talk about linking across a number of administrative data sets but not how this was done. b. What are the implications of the lack of a unique numeric identifier for building longitudinal profiles of individuals' contacts with the system with regards to repeat contacts with the systems over time. Currently it appears that the work is all cross sectional (within years). In the Strengths section they talke about near complete pathways but how can this be with 64% of identifiers missing? 2. The level of missing data. a. Some LAs did not submit data. The LA's appear to opt in and out of the data collection. b. There is no clear list of the variables that are actually available (and the level of missing data associated with each of these variables and how this changes over time). There is an interesting list of Primary needs status in the Appendix but no description about how each of these variables was operationalized. There is some mention in the limitations about variations in recording practise but not any information about the impact of these variations. c. There appears to be additional information associated with the PMRs (64% missing). Are there any place based variables (post code, geographic level) that aggregate census level data on SES could be extrapolated. Using administrative data for social science research requires a high level of understanding of the data sources, collection methods and extraction processes. The authors have tried to summarise their experiences and knowledge of these data that they have gained in extracting and cleaning these data. From my personal experience it is obvious that they have invested a lot of time in energy in this process. However I do not think this paper adequately reflects this work. As a suggestion I think the paper should be refocused with the aim to document the processes required to make the CIN valuable for researchers.

REVIEWER	Paul Bywaters University of Huddersfield
REVIEW RETURNED	09-Aug-2018

GENERAL COMMENTS	While the CIN is potentially a valuable dataset for more extensive research use, it is a well known resource. It is unclear who this introduction is aimed at and whether it sufficiently and accurately outlines the strengths and weaknesses of CIN data. The abstract suggests that the CIN dataset holds 'longitudinal' data. However, it is a series of annual snapshots as is acknowledged later in the paper. The CLA dataset is longitudinal. The article identifies some of the (severe) problems in creating a

	longitudinal data set from the annual CIN records but by no means all. Paragraph 2 rightly identifies that children looked after (CLA) are included in the definition of children in need. However the CIN data set is separate from the CLA data set held by the Department of Education. This should be made clearer from the outset. No data on CLA can be secured directly from the CIN dataset. The limitations of NPD data linkage should be acknowledged. For example, the NPD includes only some children of school age with large gaps for pre-school children and young people over minimum school leaving age, as well as gaps in the school age population, such as children at independent schools. While the paper acknowledges that a single episode may be recorded multiple times over successive years, it is less clear that the same child may be referred several times within or over years as well as in different LAs. It is therefore unclear whether and how this has been accounted for in calculating the total number of children in the period 2008-16. This affects other assertions of the paper. For example, it is said that 10.7% of children proceeded to a child protection plan. It is unclear whether this means 10.7% of each referral episode or 10.7% of children in need, or 10.7% of children in England. The poor quality of the disability data is not recognised. The proportion of children said to be disabled in different local authorities varies wildly, a product of radically different approaches to identifying and responding to disability. The data cannot be used as a reliable valid record of disabled children. Under findings to date, a published paper is said to have shown that more affluent LAs have higher rates of child welfare interventions. This is an inaccurate account of the paper. The reference to findings on educational outcomes is also over-simplistic. The paper fails to recognise the problems in data linkage to free school meals. In reporting national data, the paper fails to recognise the high degree of variation in patterns of CIN and CLA between local authorities.

REVIEWER	Calum Webb Research Associate, Department of Sociological Studies (Sociology, Social Policy, & Social Work) The University of Sheffield, UK
REVIEW RETURNED	24-Aug-2018

GENERAL COMMENTS	This is a valuable profile of the use and potentials of episode-level CIN census data which clearly outlines the processes involved in properly creating unique identifiers for children and episodes. Features of the data is also outlined very clearly, with careful consideration to the limitations. The coverage of the strengths and weaknesses of the data is extensive, but may somewhat downplay some of the limitations. I have appended a list of minor points that I hope include enough detail to warrant their consideration in a revision of the manuscript. 1) Limitations a) Children who are adopted or those at significant risk are issued new unique pupil numbers and, if there are significant risk factors, their records of older UPNs are deleted, which may make tracking

outcomes for these children with linkage to school data difficult (see 6.5, 6.6: https://assets.publishing.service.gov.uk/government/ uploads/ system/uploads/attachment_data/file/668524/UPN_ Guide.pdf). In the paper (p10, Unique Identifiers) you mention that adopted children receive a new ID, but I read this as referring to a new LA ID, rather than a new UPN. The change in UPNs may have some consequences for the conclusions drawn here around matching PMRs. I think this should be addressed in the paper.

b) As well as variations in recording practices, local authorities operate on varying definitions of thresholds for CIN, CPP, LAC procedures (Section 31 of the Children's Act 1989; further discussed in Bywaters, et al. 2014: doi:10.1111/cfs.12154), therefore users of the CIN data should be aware that the definition of a vulnerable child may differ between LAs. I think this point is alluded do but could be developed further in the profile. I think this should be addressed somewhere in the paper.

c) Although it is possible to link environmental information (such as that about area deprivation), doing so considerably complicates the ethical considerations. Since cases of child abuse and neglect, CPPs, etc, are relatively small within areas this introduces a risk of identification of vulnerable children and families if researchers wish to, for example, obtain data on LSOAs/postcode areas with which to link area-level socioeconomic information. This would likely introduce a number of additional safeguards that the authors may want to make public health, epidemiology, geography, or sociology researchers aware of. I don't think this needs to be in the paper, but I feel like it would be a helpful addition.

2) Strengths

a) I think some of the claims in the strengths section about capturing children at the edge of social services risk being unfounded and that the wording could be softened. Certainly it offers some potential to do this, but the problems of this kind of data could be addressed. Literature in Social Work has often found that under high workload pressures, especially in local authorities with large numbers of vulnerable children, social work professionals can often employ coping mechanisms to deflect contacts. This is documented in work by Broadhurst, et al. (2010) 'Performing Initial Assessment' doi.org/10.1093/bjsw/bcn162 ; I think the paper would benefit from the inclusion of this proviso.

3) Data overview

a) I wonder if the authors might see the benefit of investigating the patterns of missingness in the data, and whether missing data appears to be missing completely at random, missing at random, or not missing at random? There may be consequences for researchers using the dataset if missing data is not missing at random and listwise deletion is used to handle this. These consequences are covered well in Little, et al., 2013: doi.org/10.1093/jpepsy/jst048 I think this would be a helpful addition for the target audience of the paper and for the overview of the data itself, but I accept that it may not necessarily fit within the scope of the paper, so leave any action on this comment up to the discretion of the authors.

	4) Miscellaneous a) There is a misinterpretation of the findings in a paper by Bywaters, et al. (bibliography 17) on page 15 (lines 35-40). Bywaters, et al., (2018) found that although more deprived local authorities had higher rates of CPP and LAC than less deprived local authorities, children living in comparably deprived neighbourhood areas in less deprived local authorities were more likely to be on CPP or LAC than those in more deprived local authorities (the 'Inverse Intervention Law'). There is a short video on the Child Welfare Inequalities information page explaining this that may be of help: https://www.coventry.ac.uk/research/research-directories/current-projects/2014/child-welfare-inequality-uk/ This should be addressed before publication. b) Page 25, line 25, there is a misplaced comma in the total LA child IDs.

VERSION 1 – AUTHOR RESPONSE

We are very grateful for the thoughtful feedback from reviewers, and have incorporated their suggestions into our manuscript. Please see our response to reviewer comments below. Where similar comments were made between reviewers, we have summarised them and addressed them together to ensure our response is concise.

1. **Lack of clarity around the purpose of the paper (Reviewer 1 and 2)**
   We agree with reviewers. We have amended the following sections to clarify the purpose of the paper. This diverges slightly from the BMJ Open Cohort Profile guidance, but we believe this significantly improves the paper & request the Editors' discretion:
   a. We have re-written the **abstract**, to focus on the purpose of the paper rather than the Children in Need Dataset. It now provides a summary of the participants and data instead of 'findings to date' and 'future plans.'
   b. We have reworded the **article summary** to "strengths and limitations of the Children in Need Census (CIN)."
   c. We have re-written the **introduction**, clarifying the purpose of the paper as well as elaborating on the CIN dataset (see below).

2. **Further clarification around the CIN population (Reviewer 1, 2, and 3)**
   We agree with the reviewers, and have provided further context around CIN. We have provided the following information on the CIN population:
   a. We have re-written the **introduction** with readers in mind who are unfamiliar with English Local Authorities, and have provided further details.
   b. Within the **Background to the Children in Need Census** we have included the % of children in care and % children on child protection plans as a proportion of children identified as in need, which provides further context for the CIN population.
   c. In **Figure 1**, we have included the estimated population sizes based on DfE's statistical releases for 2015/16, including the number of children attending a publicly funded school (i.e., found in the School Census). We hope this clarifies the relationship between CIN and other datasets in the NPD.
   d. We are wary of providing further information on the School Census as this is an entirely different dataset, and we feel the relevant information for our paper is the % of children who can be found in the School Census rather than characteristics of the School Census or the variables therein. While we recognise that the School Census may itself benefit from a Cohort Profile, we believe this is beyond the scope of the current paper. Under **Data available in CIN**, we have highlighted that further information on the School Census and CLA is available in the NPD user guide, data collection specification, and data tables.
   e. Reviewer 2 expressed some uncertainties around whether our descriptive statistics of the CIN data was % of children or % episodes. Note, we specify whether each value is attributed to an episode or a child. Nonetheless, we have clarified the following: Under

Unique Identifiers, we have reworded it to "In our data, 36% of children did not have *any record of* a PMR…" Under **Case Information,** we have made it explicit that the presented statistics on primary need statuses are based on the first available referral information for each child, and have added that 10.7% of children *in our data* went onto a child protection plan *at some point*. We hope these changes make it explicit that our descriptive statistics are to do with children rather than episodes.

3. **Further clarification of drawbacks/issue in the data (Reviewer 1, 2, and 3)**
We agree that further elaboration is required around issues with the data. Some of the requested clarifications were provided/quantified in our supplementary material. We have made this more explicit in the main text, and have made the following amendments:
    a. **Differences between LAs (Reviewer 2 & 3):** Under **Background to the Children in Need Census** we have made it explicit that the thresholds around children in need vary between local authorities. We have also clarified that referrals for needs assessment must first be 'accepted' by LAs for it to be recorded in the CIN census. We have also changed **the title** to 'records of children referred for social care support in England' (i.e., away from vulnerability) which is a more accurate reflection of the content of CIN, and have made it explicit that the legal definition of 'in need' is open to professional interpretation/varying area-based thresholds in the **introduction**. We believe this implicitly and explicitly outlines that the thresholds around 'children in need' varies between LAs.
    b. **Ethics/Risks (Reviewer 3):** Under **Data Access** we have elaborated on the ethics and risk of data usage, particularly around risk of identification of vulnerable children and families. In relation to this, we have elaborated on the approval process for access to CIN. (Note that the process has changed significantly since we originally submitted the manuscript. Our revised manuscript is contains up-to-date information.)
    c. **Missing data (Reviewer 1 & 3):** Information on % of missing is provided in detail in the supplementary material, and we have clarified this in different sections under **Data available in CIN**. We have not carried out analyses on whether missing information is missing at random, as whether data is MAR/MCAR/MNAR is only relevant to specific analyses rather than the data itself.
    Note, LAs do not opt in and out of data collection. In early years of the census, some LAs (around 2 per year) were unable to submit data for various reasons (such as changing their internal IT system). Data submission to the Department for Education is a statutory requirement. Non-submission is elaborated further in the supplementary information, but we have clarified the extent of non-submission under **Data quality**.
    d. **Available variables (Reviewer 1):** The list of available variables are outlined in **Table 2**, and % missing for the variables we have access to are outlined in the **supplementary information**. We have made this more explicit in various sections under **Data available in CIN**, and we also direct readers to the variable list available from the Department for Education.
    e. **Meaning of variables (how variables are operationalised, Reviewer 1; issues with disability data, Reviewer 2):** We have clarified that the coding criteria for primary need status and disability status can be found in the collection guide for local authorities. We have also referenced a report outlining variations in how disability is recorded and coded between local authorities.

4. **Further clarifications specifically to do with linkage with other datasets (Reviewer 1, 2 and 3)**
We acknowledge that information on ID and linkage needed more clarity. We note some confusion around how CIN relates to other datasets (i.e., the school census is a completely different dataset with some shared IDs – so not under the scope of our CIN data profile). To address these issues, we have amended the following:
    a. **Children's LA Child IDs and PMRs (Reviewer 1, 2 & 3):** Under **Data available in CIN: Unique Identifiers** and **Limitations**, we have clarified that children receive a new UPN/PMR as well as new child ID at local-authority level when they are adopted. We have provided further information on PMRs, including which children have or don't have PMRs, and how this can be used for linkage with the School Census. We have elaborated on the limitations of the different IDs, specifically around tracking children

through time. We have also added some contextual information on how many children were adopted from care in 2015/16.

b. **Other possible identifiers (Reviewer 1):** We clarify that geographical information smaller than local authorities are available, but access to this information is severely restricted due to risk of identification and sensitivity of data. We have clarified this under **Data available in CIN: Children's Characteristics**. We also refer to issues around accessing identifiable information under **Data Access**.

c. **Linkage with other data sources (Reviewer 1 & 2):** We hope additional information on population size in **Figure 1** clarifies how the CIN census relates to other datasets within the NPD. Our elaboration on Child IDs above clarifies how children can be linked to the School Census (if they have PMRs).

5. **CIN is already a well-known resource (Reviewer 2)**
We do not agree that the CIN is well-known. In our experience, knowledge of CIN is limited to those who work with/research Children's Services in England. The lack of metadata means it is difficult for new users to understand the dataset, and disadvantages those who do not have personal connections to individuals who have previously used CIN. We have received requests for pre-prints, and have fed back to the Department of Education with our data description. We have clarified the need for a data description in the **introduction**, with the target audience being researchers who are unfamiliar with the English children's social care system.

6. **CIN is not a longitudinal dataset (Reviewer 2)**
We do not agree: CIN holds longitudinal data as dates of events are provided, and researchers are able to follow children/cases through time. Most longitudinal datasets are constructed from snapshots, and can exist in different structures (long, wide). We therefore keep the description of CIN as longitudinal, but clarify that the dataset is in long format with repeated entries under **Data Structure**.

   We do not provide specific instructions on how to restructure data, as the appropriate method depends on how the researcher intends to use the data. Note, we have provided a verbal outline of data cleaning that is required in the supplementary material. We have made the reference to this guidance more explicit under **Data Structure**.

7. **Reframe the paper to focus on processing of data (Reviewer 1)**
While we agree that this would be very useful for researchers, the central purpose of the Cohort Profile is to provide a brief introduction to the dataset. Therefore, we believe structuring the paper around processing CIN will not meet the purpose of a cohort profile. Instead, we have expanded our verbal description of our CIN cleaning methods in our **Supplementary Information**, and have invited readers to contact the corresponding author for more information.

   We are wary of providing specific instructions around data processing as this depends on what researchers want to do. Researchers also receive tailored 'CIN extracts' created by DfE, so may not necessarily correspond to our copy of CIN. Nonetheless, we hope our Cohort Profile will be beneficial for researchers so they are able to plan ahead.

8. **Misinterpretation of CIN findings (Reviewer 2 & 3)**
We thank the reviewers for spotting this. We have amended and elaborated on the findings by Bywaters et al., as well as Sebba et al. under **Findings to Date**.

Additional amendments:

- Since submission, the process of accessing CIN has been amended by the Department for Education. We have updated the access methods under **Data Access**.
- We have amended some typos in the manuscript.
- We have included a **Public Involvement** section to meet the new BMJ Open requirement of outlining participant and public involvement in studies.

VERSION 2 – REVIEW

REVIEWER	Anna Stewart Griffith University, Australia
REVIEW RETURNED	01-Nov-2018

GENERAL COMMENTS	I enjoyed reading this revised version of the paper. I learnt alot about how the system works in the UK, how the data are collected and the strengths and weaknesses associated with using these data for research. I am sure that researchers who are interested in using this data repository will find this paper a valuable resource. I have a couple of minor points. On page 15 you described the children's characteristics. You state that 2,182 children were recorded as intersex or other. However the other information about gender was reported as percentages, and add up to 99%. Does this mean that 2,182 children is 1%. Can you clarify this please? Also the paper seems to come to an end very abruptly. It would be nice to see some sort of concluding statement.

REVIEWER	Paul Bywaters Huddersfield University
REVIEW RETURNED	29-Oct-2018

GENERAL COMMENTS	This paper is of potential value to researchers with little knowledge of the Children in Need dataset. It might be more valuable if a similar analysis was conducted of the overlapping Children Looked After Data, but that was not the task you set yourselves. It is clearly written and outlines many of the difficulties in using the data set. I would like to see a more critical approach to the quality and value of the data. 1. The child disability data is treated as if it were relatively unproblematic. There is a comment on variability of recording between LAs on page 16 but this does not go far enough in questioning the validity and reliability of the data. It is suggested that this will be discussed in the section on Limitations but I could see no further reference. 2. The data on primary need categories does outline the fact that only one need category is permitted but does not acknowledge that need categories may be used as a bid for resources by front line staff (child protection cases will be prioritised) rather than as an accurate reflection of family difficulties. It would be helpful to underline the limited pre-set categories that can be ticked - for example, the absence of any categories that relate to family socio-economic circumstances or to the role of factors like domestic violence, present in over 50% of all assessed cases. 3. I don't think it is right to say that lower level geographical data is stored in CIN (p.16). It will be stored by LAs but it is not reported or stored centrally. It can be linked through PMR for school age children but reflects placement data rather than home circumstances for CLA. 4. In Table 1 you record some very large (>50%) year on year rises and falls in numbers of recorded children and episodes. I think these deserve examination as to whether the data are reliable and valid or whether they reflect changes in the realities of families' lives, in front line practice, in recording practice or in the

	data recording guidance. 5. On page 21 you summarise research as showing 'that LAs with overall lower levels of deprivations were more likely to intervene on families living in the more deprived neighbourhoods than LAs with overall higher levels of deprivation'. The research actually says that low average deprivation LAs intervene more at all levels of neighbourhood deprivation than high average deprivation LAs. 6. When discussing the ethnicity of the children you give the percentages without adjusting for the 6% missing data. I think the proportions by ethnic group should be based on the data where ethnicity is available. 7. I am not sure that the calculation of 2.7 million children is accurate. I think you methodology allows you to talk about 2.7m child IDs but you cannot know - I think - how many children will have moved to another LA and given a separate ID when a new need episode starts or been adopted and changed ID. I am open to persuasion on this but I cannot see how you can be sure that this number reflects children rather than separate IDs. This would affect the Abstract as well as the main text.

REVIEWER	Calum Webb The University of Sheffield, United Kingdom
REVIEW RETURNED	14-Nov-2018

GENERAL COMMENTS	I am grateful to the authors' thoughtful responses to the comments made in the first round of review and feel like the changes that have been made sufficiently address my original review. I recommend this manuscript for publication and think it will be a valuable resource for researchers in the future.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

1) I have a couple of minor points. On page 15 you described the children's characteristics. You state that 2,182 children were recorded as intersex or other. However the other information about gender was reported as percentages, and add up to 99%. Does this mean that 2,182 children is 1%. Can you clarify this please?
This was a consequences of using integers. We have changed the percentages to 1dp, so it adds up to 100%.

2) Also the paper seems to come to an end very abruptly. It would be nice to see some sort of concluding statement.
We have introduced a short paragraph under *Limitations* reflecting on further research which is required for a better understanding of the CIN, and an additional short paragraph under *Data Access* to link back to the introduction.

Reviewer: 2

1) The child disability data is treated as if it were relatively unproblematic. There is a comment on variability of recording between LAs on page 16 but this does not go far enough in questioning the validity and reliability of the data. It is suggested that this will be discussed in the section on Limitations but I could see no further reference.
Under *Children's Characteristics* and *Limitations*, we now explicitly mention disability data may be particularly susceptible to validity and reliability issues. We have also provided additional references to research papers which elaborate on this issue in *Limitations*, and reiterate the difficulties around interpretation of administrative data.

Note, as a descriptive paper (rather than analytical), our aim is to flag known issues rather than provide an in-depth critique of each variable. We recognise the trade-off here is that, for researchers who hold specialist knowledge of CIN, it may seem like our description does not capture enough of the complexity behind CIN. Our intention is not to dismiss data issues around disability data. Rather, we believe more detailed critique of specific aspects of CIN would be better served by analytical research papers, which we guide readers to where they are available. (In fact, we are currently working on a mixed-method paper analysing the validity and reliability of referral information in CIN- so we do fully recognise this issue.)

We believe this addresses the valid suggestion around highlighting limitations around disability information in CIN, without crossing over into a critical analysis of specific aspects of CIN.

2) The data on primary need categories does outline the fact that only one need category is permitted but does not acknowledge that need categories may be used as a bid for resources by front line staff (child protection cases will be prioritised) rather than as an accurate reflection of family difficulties. It would be helpful to underline the limited pre-set categories that can be ticked - for example, the absence of any categories that relate to family socio-economic circumstances or to the role of factors like domestic violence, present in over 50% of all assessed cases.

Under *Case Information* we have clarified that primary need status may not reflect actual need. Under *Limitations* we have elaborated that different incentives may exist around different primary need categories. Specific need categories are outlined in the supplementary information, and we refer readers to LA guidance with further details. As above, we hope our elaboration on difficulties in interpreting administrative data clarifies possible validity and reliability issues.

3) I don't think it is right to say that lower level geographical data is stored in CIN (p.16). It will be stored by LAs but it is not reported or stored centrally. It can be linked through PMR for school age children but reflects placement data rather than home circumstances for CLA.
We agree this was unintentionally misleading, as this information is not available for all episodes. We have clarified that the lower level geographical data is available for episode between 2008 and 2010 only. This information is held by DfE.

4) In Table 1 you record some very large (>50%) year on year rises and falls in numbers of recorded children and episodes. I think these deserve examination as to whether the data are reliable and valid or whether they reflect changes in the realities of families' lives, in front line practice, in recording practice or in the data recording guidance.
We have made explicit in Table 1 that episode numbers in 2008/09 are much fewer as the census period started in October instead of April (i.e., 6 month period instead of 12 months).

Regarding the jump in number between 2011/12 and 2012/13, at present we are not sure what caused the increase in the number of children in CIN. Archived public documents show that a review of CIN was being undertaken in 2010/12, which led to some changes in the data being collected as well as the data submission process. This period also overlaps with the publication of the Munro Review, which could have influenced the number of referrals as well as data recording practices. We have introduced a comment on this under *Data Collection Process*.

5) On page 21 you summarise research as showing 'that LAs with overall lower levels of deprivations were more likely to intervene on families living in the more deprived neighbourhoods than LAs with overall higher levels of deprivation'. The research actually says that low average deprivation LAs intervene more at all levels of neighbourhood deprivation than high average deprivation LAs.
Our interpretation is that there is an interaction effect between local and LA-level deprivation. We have amended this to clarify intervention is higher at all levels of neighbourhood deprivation.

6) When discussing the ethnicity of the children you give the percentages without adjusting for the 6% missing data. I think the proportions by ethnic group should be based on the data where ethnicity is available.

We disagree with this as our aim is to describe CIN as a whole, rather than children whose ethnicities have been recorded. Missing data can be important data in itself, depending on the research question- so we are reluctant to remove cases from the descriptive statistics where ethnicity is not available.

7) I am not sure that the calculation of 2.7 million children is accurate. I think you methodology allows you to talk about 2.7m child IDs but you cannot know - I think - how many children will have moved to another LA and given a separate ID when a new need episode starts or been adopted and changed ID. I am open to persuasion on this but I cannot see how you can be sure that this number reflects children rather than separate IDs. This would affect the Abstract as well as the main text.

We have slightly changed the wording in the abstract, article summary and main text to clarify this is an estimate. Note, talking about 2.76m *Child IDs* may also not be precise – for example, children with multiple LA IDs have been counted as 1 if we are able to track them between LAs through their PMRs.

We note that, technically, the 2.76m is to do with *derived child IDs*. However, we feel it would be difficult for readers to follow if we did not talk about "children" and replaced this with "derived child IDs" - particularly as there are multiple ID variables in CIN.

While the 2.7m value is not a *precise* number of children, we believe it is an *accurate representation*: Given that IDs always serve as proxies of a child in a child-based dataset (acknowledging that with proxies there is always risk of error), the number of derived child IDs is the best estimate for the number of children in CIN. Our paper puts substantial focus on the issues around child IDs & how there are no truly unique identifiers (one of the main challenges in CIN). We explain throughout the manuscript and SI, both explicitly and implicitly, that 2.76m is estimated. We are therefore confident that readers will understand this is an estimated figure, and we do not claim precision.

Reviewer 3

Thank you for your positive assessment of our paper and its importance.

Other amends:

We have included a Future Plans section in the abstract, outlining future data collections and data availability. This information has also been added under Data Access.

Further investigation through archived public documents have revealed that CIN was submitted through the COLLECT system prior to 2012/13. We have therefore removed our statement about uncertainty in the data collection process prior to this date.

We have removed our reference to the Children's Social Care (CSC) Data User Group. There have been some delays and unfortunately the site is currently not ready to go live.

VERSION 3 – REVIEW

REVIEWER	Paul Bywaters Huddersfield University, UK
REVIEW RETURNED	12-Dec-2018

GENERAL COMMENTS	This version of the paper answers concerns raised previously and provides a valuable introduction to CIN data.