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 Copenhagen Primary Care Laboratory Pregnancy (CopPreg) Database
AUTHORS	Janbek, Janet; Kriegbaum, Margit; Grand, Mia; Specht, I.; Lind, Bent; Andersen, Christen; Heitmann, Berit

VERSION 1 – REVIEW

REVIEWER	Claire Thorne University College London, UK
REVIEW RETURNED	02-Oct-2019

GENERAL COMMENTS	This cohort profile manuscript presents a very valuable data resource – the CopPreg database, which includes 203,608 pregnant women who delivered in the Copenhagen area between 2000 and 2016, and their 348,248 children; unusually, the database also includes data on some of these children's fathers. The CopPreg database has been created through linkage between the pre-existing Copenhagen Primary Care Laboratory database and the Danish Medical Birth Register, and this paper focuses on the sub-groups within the database with laboratory test data available. Some specific comments are given below: The abstract could be strengthened by some additional clarity regarding the clinical test requisitions (i.e. those conducted in primary care only). Page 1 The strengths of the cohort should also include the potential for linkage with other Danish national registries (as outlined later in the manuscript) Regarding women who were referred to hospitals, would linkage with the NPR allow access to data on lab / clinical tests conducted in hospital (out- or in-patient)? (page 1 and page 6) Page 2 – line 60 – "to examine pre- and antenatal risk factors.." – please edit / explain as prenatal and antenatal mean the same? Should this say "preconception"? Page 4- Please provide a definition of stillbirth (line 97). From a later footnote, it seems that this definition has changed over time. Page 4 – It would be helpful to understand a bit more context regarding general antenatal care in Denmark – particularly, more information on the maternity care level score and what the different levels mean (maybe with some examples). Page 7 – It is stated that statistical analyses can be conducted at aggregate level only. Please could this point be expanded or confirmed, as the lack of availability of anonymised individual patient-level data is important, as this precludes some specific analytic approaches. It also seems to contradict information on

	page 14, where data linkage on an individual level with other registers is discussed. Page 9 and Table 1 – please clarify regarding the definition of parity. In the text it states that most women "previously had a total of 1-2 children", which implies parity at the time of the pregnancy of interest in the cohort – but from the Table, there are no nulliparous women, so the text seems to be referring to parity after the index pregnancy in the database. Table 1 – there are two footnotes with a * Table 2 – "Gestational age in days" should be relabelled, as this is not what is being presented There is a high level of detail provided concerning the laboratory measures available, with a key strength being that these tests were all conducted in the same laboratory, with information on reference limits available. A weakness is that point of care tests conducted at GP clinics were not included, but the authors acknowledge this. Page 15 – The section on data access should be expanded to give more information on processes and governance. Trivial points / typos - "visits" instead of "visitations" (page 6) - Line 359 – "enabling" - Line 364 0 "by the GPs FOR an unknown indication"

REVIEWER	Ibrahim Hammad Intermountain Healthcare and University of Utah Health, Salt Lake City, UT, USA
REVIEW RETURNED	27-Oct-2019

GENERAL COMMENTS	I congratulate the authors on carrying out what was a meticulous and thorough descriptive paper. They've described with great detail the structure and components of The Copenhagen Primary Care Laboratory Pregnancy database. The database is the result of a linkage of two databases, a birth registry, and laboratory database. With that being said, I have concerns about the need to publish this paper. Multiple databases have some similarities with this database. To count a few, The Perinatal Database of the Netherlands, the Finnish Drugs and Pregnancy database, and the Utah Population Database. The concept is not novel. Even if this particular database does have some advantages including laboratory data on the fathers or the ability to follow individuals over a long period. The database is a rich source for future studies, as noted by the authors. The description of the database can be summarized in the material and method section. Additional information with regards to the methodology can be provided as a supplement if needed.

REVIEWER	James M Roberts Magee-Womens Research Insitute Department of Obstetircs Gynecology and Reproductiev Sciences, Epidemiology and Clinical and Translational Research University of Pittsburgh
REVIEW RETURNED	28-Oct-2019

GENERAL COMMENTS

This presentation describes a remarkable resource consisting of laboratory studies, and clinical outcome on pregnant women, their husbands, and their children. The linking of several databases provides an enormous amount of data which has the potential to be linked to several other Danish databases. Although the data is extensive there are important limitations. Many of these are mentioned some are not. In those mentioned, it is very important to more strongly emphasize these limitations. Finally, many apparent limitations are potentially modifiable with information that could be made available by the authors of the database. The availability of this information should also be mentioned.

The limitations include indication bias related to what laboratory studies happen to be available on the mother, baby and father. These are alluded to, but you should provide more emphasis. Perhaps the most interesting findings related to inflammation in cardiovascular metabolic changes, are clearly collected on a specific subset of patients. There are of course studies which are done on the vast majority of patients and these of course would not be subject to bias Unfortunately, the most interesting findings are those associated with substantial bias. While multiple imputation and IPW are methods for handling missing data/indication bias, these techniques are difficult to implement when more than half of the data is missing as is the case for many of the variables in this database. In theory, missing data can be handled, but the discussion section oversimplifies this concern and even states that the current investigators have not resolved this issue in lines 366 and 367. Without, a clear solution or approach to handling missing data, it is difficult to see how any researcher using this dataset will be able to provide results that are not biased. In addition, the gestational age when the studies are obtained is of course relevant and although this information is discernible this should be mentioned as a limitation which can be overcome. Also, some of the most interesting data, information obtained before delivery is not described.

There are several specific modifications that could you improve the presentation:
1. In table 2 there is no description of small for gestational age infants. It is of course obvious that with the availability of gestational age birth weight infant sex it is possible to calculate this data, however the inclusion of low birth weight which is a combination of preterm birth and small for gestational age infants it does not seem useful.
2. On page 14 “creatininium” should probably be “creatinine”.
3. In figure 1 on page 27, the arrows make it unclear how the final number of people in the database was reached. It is difficult to understand what the arrow from the box with N= 608,898 women to the box with N=1,061,130 represents. Likewise, the arrow leading to the final CopPreg Database box is not interpretable. Can this be clarified?
4. On page 29, I appreciate the difficulty of trying to provide all the information that is included in figure 3. However, several aspects of the figure are quite confusing. Why are there arrows connecting pregnancies and fathers and then pregnancies and fathers with children? Symbol A presents that of 105,447 of pregnancies with requisitions have fathers who also have requisitions. However, in B it stated that 39,815 fathers had periconception requisitions. And with symbol C a similar discrepancy exists in terms of children with

	requisitions with 65,315 in C and 42,492 in A. Can this be clarified? 5. The gestational age when a particular test is obtained in pregnancy is quite relevant. I suspect this is too complicated to include in the paper, but it should be made clear that this information is available.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1
Reviewer Name: Claire Thorne
Institution and Country: University College London, UK Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below.

This cohort profile manuscript presents a very valuable data resource – the CopPreg database, which includes 203,608 pregnant women who delivered in the Copenhagen area between 2000 and 2016, and their 348,248 children; unusually, the database also includes data on some of these children's fathers. The CopPreg database has been created through linkage between the pre-existing Copenhagen Primary Care Laboratory database and the Danish Medical Birth Register, and this paper focuses on the sub-groups within the database with laboratory test data available.

Reply:

Thank you for taking the time to review our manuscript. We highly appreciate it. Many thanks for bringing our attention to some important issues to clarify our points better and communicate our findings. We have now adjusted the manuscript accordingly. Below are some comments to the specific points made.

Some specific comments are given below:

The abstract could be strengthened by some additional clarity regarding the clinical test requisitions (i.e. those conducted in primary care only).

Reply:

Thank you, we have now added in lines 15-16 in the abstract (in the marked copy of the main document) emphasis that these tests were ordered by GPs in the primary care setting only.

Page 1 The strengths of the cohort should also include the potential for linkage with other Danish national registries (as outlined later in the manuscript)

Reply:

Thank you, you are right, we overlooked mentioning this in our summary of strengths. We have now added a brief sentence stating the potential for linkage in lines 32-33 in the marked copy of the main document.

Regarding women who were referred to hospitals, would linkage with the NPR allow access to data on lab / clinical tests conducted in hospital (out- or in-patient)? (page 1 and page 6)

Reply:

Thank you for this very relevant comment. Technically it is possible to link to hospital (in- and out-patient) laboratory data, but the challenge is often the lack of knowledge about the data you link to. There is a national laboratory database in Denmark with hospital data. However, it does not cover the whole CopPreg database period (only 2010-2015) and it has several limitations e.g. several laboratories contribute with data and the differences in test results compared to CopPreg data may be the result of different preanalytical, analytical and postanalytical methods and procedures used by the two laboratories rather than true biological differences. The limitations may not restrict the use of the results in patient treatment but may have undesirable consequences when used in research (comparability of test results over time, selection bias etc.). We have further included this information in the manuscript (Lines 425-434 in the marked copy of the main document) as a possible limitation.

Page 2 – line 60 – “to examine pre- and antenatal risk factors..” – please edit / explain as prenatal and antenatal mean the same? Should this say “preconception”?

Reply:

True, we have now modified in the text to prenatal only, as we intended to point out risk factors during the entire prenatal period (before birth, around pregnancy). (Line 67 in the marked copy of the main document)

Page 4- Please provide a definition of stillbirth (line 97). From a later footnote, it seems that this definition has changed over time.

Reply:

We have now added a definition of stillbirths in parenthesis here also (lines 112-115 in the marked copy of the main document).

Page 4 – It would be helpful to understand a bit more context regarding general antenatal care in Denmark – particularly, more information on the maternity care level score and what the different levels mean (maybe with some examples).

Reply:

Thank you for this. We have now added some more examples regarding the maternity care level score and the care offers a woman receives (lines 142-146 in the marked copy of the main document). We hope this provides more information needed to gain an overview of Danish antenatal care system.

Page 7 – It is stated that statistical analyses can be conducted at aggregate level only. Please could this point be expanded or confirmed, as the lack of availability of anonymised individual patient-level data is important, as this precludes some specific analytic approaches. It also seems to contradict information on page 14, where data linkage on an individual level with other registers is discussed.

Reply:

Many thanks for bringing our attention to this. We of course mean individual level linkage. Choice of phrasing as aggregate level was to describe that no persons can be identified, and that results can only be reported at an aggregated level. We have modified in the manuscript lines 196-197 in the marked copy of the main document. We hope it is better this way.

Page 9 and Table 1 – please clarify regarding the definition of parity. In the text it states that most women "previously had a total of 1-2 children", which implies parity at the time of the pregnancy of interest in the cohort – but from the Table, there are no nulliparous women, so the text seems to be referring to parity after the index pregnancy in the database.

Reply:

Thank you for spotting this. Yes, the reason there are no nulliparous women is because the index pregnancy was counted in and as such parity is at least 1 for all women. Women with parity of 2 and above refer to the number of births/deliveries for each woman before the index pregnancy in the cohort. We have clarified this in text now in lines 263-264 in the marked copy of the main document.

Table 1 – there are two footnotes with a *

Reply:

Thank you for bringing attention to this, we have now removed the first * and kept it as a footnote only since it is meant to comment on the entire table. We kept the second one which refers to BMI in the table.

Table 2 – "Gestational age in days" should be relabelled, as this is not what is being presented

Reply:

Thank you, we agree. We have now deleted the 'in days'.

There is a high level of detail provided concerning the laboratory measures available, with a key strength being that these tests were all conducted in the same laboratory, with information on reference limits available. A weakness is that point of care tests conducted at GP clinics were not included, but the authors acknowledge this.

Reply:

Thank you. We agree.

Page 15 – The section on data access should be expanded to give more information on processes and governance.

Reply:

We have now added more information regarding how data is handled and how such data can be accessible to researchers in lines 370-373 in the marked copy of the main document, thank you!

- Trivial points / typos
- "visits" instead of "visitations" (page 6)
- Line 359 – "enabling"

- Line 364 0 “by the GPs FOR an unknown indication”

Reply:

Thank you, we have now fixed these and looked more thoroughly throughout modifying text in other places as well.

Reviewer: 2
Reviewer Name: Ibrahim Hammad
Institution and Country: Intermountain Healthcare and University of Utah Health, Salt Lake City, UT, USA Please state any competing interests or state ‘None declared’: None

Please leave your comments for the authors below.

I congratulate the authors on carrying out what was a meticulous and thorough descriptive paper. They've described with great detail the structure and components of The Copenhagen Primary Care Laboratory Pregnancy database. The database is the result of a linkage of two databases, a birth registry, and laboratory database. With that being said, I have concerns about the need to publish this paper. Multiple databases have some similarities with this database. To count a few, The Perinatal Database of the Netherlands, the Finnish Drugs and Pregnancy database, and the Utah Population Database. The concept is not novel. Even if this particular database does have some advantages including laboratory data on the fathers or the ability to follow individuals over a long period. The database is a rich source for future studies, as noted by the authors. The description of the database can be summarized in the material and method section. Additional information with regards to the methodology can be provided as a supplement if needed.

Reply:

Thank you very much for taking the time to review our manuscript. We appreciate your comments and your acknowledgment of the work that has been put into the manuscript and establishment of the database. There are of course other laboratory databases available, each with their own strengths and limitations, naturally. As you kindly point out, our database is in fact unique, and remains unique when compared with other available databases such as those pointed out, with regards to the very large amount of clinical test results available for fathers (in addition to pregnant women and children), while at the same time, all health and disease as well as socio-demographic information are readily available for the fathers themselves but also for their children and the related to pregnancies. We believe that such data resource is absent in its sample size, coverage of data on mothers, fathers and children and perhaps most importantly, in the fact that data is readily available and already analyzed where no additional costs in relation to analyzing any samples or collecting additional data, are needed. We also believe that longitudinal follow up of our population is in fact a big advantage, as you also kindly mention. That said, we thank you for mentioning examples of these available databases, which we now refer to and put focus on where we believe our database has strengths for use in research (lines 71-75 in the marked copy of the main document). Thanks again.

In a time where cohorts are continuously being recruited, costs are continuously going towards data collection (and sample analysis), we believe and aspire to reach out to both the national and international community to communicate the availability of data that is highly needed in early disease programming field of research. Moreover, the availability of such a data recourse to the international community, i.e. relative ease of access and utility of such data is yet another important advantage, which we also aim to communicate. We thank you again for your comment.

Reviewer: 3
Reviewer Name: James M Roberts
Institution and Country: Magee-Womens Research Insitute Department of Obstetircs Gynecology and Reproductiev Sciences, Epidemiology and Clinical and Translational Research University of Pittsburgh Please state any competing interests or state ‘None declared’: None

Please leave your comments for the authors below:

This presentation describes a remarkable resource consisting of laboratory studies, and clinical outcome on pregnant women, their husbands, and their children. The linking of several databases provides an enormous amount of data which has the potential to be linked to several other Danish databases. Although the data is extensive there are important limitations. Many of these are mentioned some are not. In those mentioned, it is very important to more strongly emphasize these limitations. Finally, many apparent limitations are potentially modifiable with information that could be made available by the authors of the database. The availability of this information should also be mentioned.

The limitations include indication bias related to what laboratory studies happen to be available on the mother, baby and father. These are alluded to, but you should provide more emphasis. Perhaps the most interesting findings related to inflammation in cardiovascular metabolic changes, are clearly collected on a specific subset of patients. There are of course studies which are done on the vast majority of patients and these of course would not be subject to bias Unfortunately, the most interesting findings are those associated with substantial bias. While multiple imputation and IPW are methods for handling missing data/indication bias, these techniques are difficult to implement when more than half of the data is missing as is the case for many of the variables in this database. In theory, missing data can be handled, but the discussion section oversimplifies this concern and even states that the current investigators have not resolved this issue in lines 366 and 367. Without, a clear solution or approach to handling missing data, it is difficult to see how any researcher using this dataset will be able to provide results that are not biased. In addition, the gestational age when the studies are obtained is of course relevant and although this information is discernible this should be mentioned as a limitation which can be overcome. Also, some of the most interesting data, information obtained before delivery is not described.

Reply:

Many thanks for taking the time to review our manuscript, and to do so in such thoroughness. We highly appreciate it and have tried to modify our manuscript according to your specific comments, which clearly strengthen the manuscript and help clarify what we want to communicate. Again, many thanks for this. We also acknowledge your point regarding emphasizing the limitations of using this database for research. We would like to emphasize more the fact that in regard to what is the most obvious limitation, indication bias, our research team and particularly a team of biostatisticians are currently working committedly on exploring this limitation by employing different statistical methods through using a case study, namely the case of prenatal vitamin D levels and childhood asthma. We describe this approach in lines 407-418 in the marked copy of the main document. We hope that this provides assurance that we are prioritizing exploration of this matter and hope to communicate our results in the near future, acknowledging as you kindly stated that without a clear solution to handling this limitation, biased results from research are difficult to avoid. It can however also be argued that for some projects we can get around the indication bias problem by simply defining the target population to be only those with a measurement as opposed to all pregnancies. Especially because this is the actual population of pregnant women that the GP encounters. So even if it turns out to be difficult to use missing data approaches to answer some research questions, there will certainly be others where the data will still be useful. Please find comments to the specific points you bring to our attention, below.

There are several specific modifications that could you improve the presentation:
1. In table 2 there is no description of small for gestational age infants. It is of course obvious that with the availability of gestational age birth weight infant sex it is possible to calculate this data, however the inclusion of low birth weight which is a combination of preterm birth and small for gestational age infants it does not seem useful.

Reply:

Thank you for this comment. It is of course essential and very important to present WGA data on our population. We have now pointed out in the manuscript the importance of this information and the possibility of obtaining it using the variables available in our database in lines 283-286 in the marked copy of the main document. We have kept the variables birth weight and gestational age in the table, presenting the what you can call "raw" data on the variables available in the database (vs. calculated variables). We thank you again for this.

2. On page 14 "creatininium" should probably be "creatinine".

Reply:

Thank you, we have fixed this now in line 345 in the marked copy of the main document.

3. In figure 1 on page 27, the arrows make it unclear how the final number of people in the database was reached. It is difficult to understand what the arrow from the box with N= 608,898 women to the box with N=1,061,130 represents. Likewise, the arrow leading to the final CopPreg Database box is not interpretable. Can this be clarified?

Reply:

Of course, thank you for commenting on this. We agree that it is confusing with the arrows and that the figure does not in fact reflect the reality of what we want to say, eg, that by merging these two registries together, the CopPreg population was identified. In this regard, we put the arrow to show that we identified our population by looking up the 608.898 women from the sample of 1.061.130 women. However, we have now modified the figure and now we do not have an arrow pointing from the 608.898 women to 1.061.130 women. Rather, we have now added further description of this process in text in lines 153-154 in the marked copy of the main document. We hope the figure is clearer now.

4. On page 29, I appreciate the difficulty of trying to provide all the information that is included in figure 3. However, several aspects of the figure are quite confusing. Why are there arrows connecting pregnancies and fathers and then pregnancies and fathers with children? Symbol A presents that of 105,447 of pregnancies with requisitions have fathers who also have requisitions. However, in B it stated that 39,815 fathers had periconception requisitions. And with symbol C a similar discrepancy exists in terms of children with requisitions with 65,315 in C and 42,492 in A. Can this be clarified?

Reply:

Thank you again for your comments. This has been very difficult to represent. We probably did not succeed with this, but we attempt again now, thanks again for bringing attention to this. We have now removed the arrows between pregnancies, fathers and children (these were meant to say that the fathers are related to the pregnancies, and the children are these mothers' and fathers' children), it is probably best to leave this out and have the three groups stand alone. Further, the 105.447 fathers and the 42.492 children are those of the pregnancies with requisitions (group A), whereas groups B and C are those of fathers and children of all pregnancies (340.891 total pregnancies identified, of which only a sub-group, A, had requisitions during 3 months preconception and during pregnancy while others had requisitions but that were outside this period). Moreover, the 105.447 fathers shown had requisitions, but were not specifically taken during a specific period (periconception period) as was the case for group B which we focused on. This is of course perhaps unnecessarily confusing, we have thus chosen to remove these 105.447 fathers and 42.492 children from the figure. These are still mentioned in text in lines 222-226 in the marked copy of the main document, but as mentioned, they are not described further in text and are not focused on. We further expand on the group of pregnancies which we did not focus on in this paper in lines 232-235 in the marked copy of the main document, to draw attention to their value although not described in detail in the paper. We really hope this gives more clarity on our population. Many thanks again for bringing attention to this.

5. The gestational age when a particular test is obtained in pregnancy is quite relevant. I suspect this is too complicated to include in the paper, but it should be made clear that this information is available.

Reply:

True, and also true that it would be complicated to include this in the paper, however possible. We have the date of each requisition/clinical test analyzed, and we have the gestation age and birth date of the child, which can be used to calculate gestational age at time of each test. Due to the large amount of tests described, this is complicated as you kindly mention. We appreciate you pointing this out and we have now added in lines 339-342 in the marked copy of the main document, about the possibility of obtaining such information for each test. Thank you!

VERSION 2 – REVIEW

REVIEWER	Claire Thorne University College London Great Ormond Street Institute of Child Health
REVIEW RETURNED	31-Jan-2020

GENERAL COMMENTS	The authors have addressed the issues raised in the review - many thanks.

REVIEWER	James M Roberts Magee-Womens Research Insitute Department of Obstetircs Gynecology and Reproductiev Sciences, Epidemiology and Clinical and Translational Research University of Pittsburgh
REVIEW RETURNED	28-Jan-2020

GENERAL COMMENTS	The authors have as best possible answered my queries and increased clarity. There were challenges in the review of their responses since the line number they gave did not match the

	appropriate lines in the marked manuscript. Nonetheless managed to find them and they did well.