Reviewer's report

Title: Infant feeding practices within a large electronic medical record database

Version: 1 Date: 18 May 2017

Reviewer: Chris Gale

Reviewer's report:

Full Title: Infant feeding practices within a large electronic medical record database

This study uses data held in a database extracted from Electronic Medical Records in Ontario, Canada, to examine rates of breastfeeding up to 6 months of age among singleton term infants and associated factors. The key findings are that EMR data were able to provide information on feeding practice for about 90% of infants in the database using automatic and hand searching techniques. Among the study babies breast feeding rates declined over time and were associated with previously described factors such as maternal age and socioeconomic status.

My main issue with this paper is that I am not sure what the main, overarching point of it was. Was it to demonstrate that EMR data held in the EMERALD database is good enough to be used for observational research in this field or was it to determine the rates of breastfeeding in this cohort and associated factors? These are quite different research questions and are not entirely compatible with each other. The background section suggests the latter aim, but the conclusion suggests the former aim.

If the aim is the latter (to describe breastfeeding), this kind of work has been done many times before and in similar many cohorts, so is not particularly ground-breaking - although it will be of local interest. If however, this paper is more about validating the use of EMR data for such research, this is potentially more interesting to me as work validating and describing such routinely-extracted data sources is lacking. I would like to see the authors focus more on one of these aims explicitly.

If the paper does seek to validate the EMRALD database for observation research in relation to infant feeding, it would be good to more clearly define a "gold-standard" against which the data are to be compared (or to discuss why there is not one available). The authors could then focus on demonstrating the data quality (data completeness, accuracy, etc.) for the EMRALD dataset in
relation to this particular question, and then examine the generalizability of EMRALD vs Ontario
more widely and test for systematic differences.

My other issues with this paper are as follows:

1. The numbers are quite small - only 8,815 infants in the EMRALD dataset vs 1.2 million
births. And only about 88% of the 8,815 infants had data on breastfeeding.

2. The authors state that the EMRALD population "were highly similar to those across
Ontario" - I disagree with this statement; there are potentially very important difference
between the EMRALD cohort and Ontario births in most of the characteristics presented
in table 1 (rural residence, maternal region of birth etc.). The way the authors have
presented the differences (Standardized difference) seems misleading to me and is not
described in detail - I would like to see a more meaningful description of the difference
such as absolute difference and CIs.

3. Given the differences between the EMRALD population and the Ontario population it
would be useful to see the authors apply an approach to account for these differences in
their results.

4. The EMRALD dataset are limited to those mother-infant dyads that visited a GP within
750 days - this is clearly not a representative group of all dyads in Ontario - this should
be explored more

5. Can the authors please be more specific about when each infant had data available and
the proportions of the total population that had data at each point (2, 4, 6 months). How
many infants had a "complete dataset" and how did they compare to the wider Ontario
population?

6. Where more than one data point was available for an infant at one time (e.g. there had
been more than 1 visit between 2 and 4 months) how did the authors decide which data
point to include?
7. I am still unclear how the authors extracted data for each point. From the description in the methods section it seems as though any visit before 60 days was used as data for "2 months" - does this mean that a visit on day 3 would be coded as 2 months? Similarly, any visit after 60 days (e.g 61 days) seems to be coded as 4 months! The authors at the very least need to provide data on the mean and SD age of the data used to describe infants at 2, 4 and 6 months.

8. The data were extracted in 2 very different ways: either automatically (4955 infants) or by hand searching (2816 infants). It is important to know how different the characteristics and results are in these two groups - there are likely to be systematic differences in both of these.

9. In the results (Table 2) there seem to be many difference factors that have been tested for significance in relation to breastfeeding at multiple different points. The authors should explain how they have dealt with multiple testing here.

10. The conclusion states that "primary care EMR data can be used to assess breast-feeding practices in large populations" - I do not feel that this conclusion is warranted from the data presented. To convince me of this statement the authors need to more clearly define the comparator against which they have examined the EMRALD database and clearly explain the data quality of EMRALD as well as considering the generalisability of EMRALD data to the larger population, qualify the systematic differences between EMRALD and the wider Ontario population and consider how these would need to be accounted for.
Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

No

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

No

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I recommend additional statistical review

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?
If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I work as part of an academic unit (Neonatal Data Analysis Unit, NDAU) at Imperial College London, which uses data held in a research database (the National Neonatal Research Database) that is extracted from United Kingdom neonatal Electronic Medical Record (EMR) entries. I do not receive any salary or reimbursement from the NDAU or the EMR providers. I am funded by grants awarded by the UK Medical Research Council (MRC) and the UK National Institute of Health Research (NIHR).

In the last 5 years I have received a travel bursary from the British Association of Perinatal Medicine to attend educational conferences, these bursaries are supported by Chiesi Ltd.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal