PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Impact of chronic health conditions and injury on school performance and health outcomes in New South Wales, Australia: a retrospective record linkage study protocol
AUTHORS	Mitchell, Rebecca; Cameron, Cate; Lystad, Reidar; Nielssen, Olav; McMaugh, Anne; Herkes, Geoffrey; Schniering, Carolyn; Hng, Tien-Ming

VERSION 1 – REVIEW

REVIEWER	Reviewer name: Sarah J Nevitt Institution and Country: University of Liverpool Competing interests: I have no competing interests
REVIEW RETURNED	28-Jun-2019

GENERAL COMMENTS	I have performed a statistical review of the manuscript "Impact of chronic health conditions and injury on school performance and health outcomes in New South Wales, Australia: a record linkage study protocol" The authors describe a data linkage study with the objective of identifying the impact of injury and chronic illness on school performance. Overall the protocol is clear and well written. I have two comments on the analysis approach and a few comments on wording Content comments 1) Record linkage: it is certainly valuable to have a third party conducted in data linkage. My comment relates to the final 'linked' dataset which will be provided to the authors for this work. I assume that the identifying information (names, date of birth) used by the third party for data linkage will not be made available to the authors of this work for protection of participant privacy? Has any consideration also been given to potentially 'sensitive' data which may be present within the databases? Particularly relating to mental health? Will any additional steps be taken either by the third party responsible for linkage or the present authors to minimise any risks of data breach or re-identification of children with potentially sensitive information disclosed? 2) Data analysis plan: there are a few areas here I am unsure on, could the authors clarify? a) Page 9, line 36: "Child injury and each chronic illness will be examined separately." So does this mean that for the primary and secondary outcomes, a separate regression model will be used for hospitalisations due to injury and due to each of the four chronic conditions?

	b) Generalized linear regression is proposed for the primary outcome of 'proportion of performances below the NMS.' Linear regression methods require continuous data to be normally distributed, consider whether this 'proportion' would be normally distributed. Perhaps using a logistic regression for NAPLAN domain score below the NMS (yes or no) – I think this is the approach described over lines 43-49. c) Related to the above comment, I suggest that hospital length of stay and hospital treatment costs are unlikely to be linear (page 10, line 25). Such data is generally very positively skewed and influenced by extreme values (longer and more costly hospital stays than expected. Consider methods which allow for the likely 'non-linearity' (i.e. skewed) nature of this information d) Page 9, line 51: Why will both relative risks and odds ratios reported? Output of logistic regression models would be odds ratios. e) Page 9, line 53: Please add further details on sensitivity analyses for potential missing data values (for example any imputation methods?) Wording comments 1) What this study adds "This study will identify the types of injuries and chronic illness associated with problems with learning at school" Unless I've missed something in this protocol, I didn't think that the 'type' of injury was considered, just whether the child had been hospitalised due to an injury or not (and injury severity from Table 2). Please clarify and maybe reword. 2) There are a few references throughout to 'years' and 'grades' of the children (e.g. page 5, objective 3; page 9, line 42). I presume these years / grades map specifically to the Australian school system? E.g. I know that the English school years are different. As the readership of the journal may be global, I suggest instead to refer to ages of children (or at least defining the ages of the children within the different school years and grades mentioned here) for clarity. 3) Page 5, line 33-34: "These five conditions were selected as…" This wording didn't quite seem right to me as the focus is on hospitalisation due to injuries or four listed chronic conditions – injuries are not a condition so it isn't really five conditions. Consider rewording 4) Page 7, line 49: "The comparison group will be randomly matched in a 1:4 ratio on age, gender and residential postcode to their matched case." How exactly are 'residential postcodes' defined here and are these areas quite broad? For example, I'm fairly confident that in the UK, it would be very restrictive to try and match age gender and exact residential postcodes – whereas if the first half of the postcode was used in the UK, this would result in a much broader sampling area. 5) Page 9, line 34: "All hospital episodes of care related to the one event will be linked to form a period of health care." I'm not completely following what 'one event' means here.

	Does this refer to all hospitalisations for the same injury, or related to one epileptic seizure, one asthma attack, one hypoglycaemic episode or one mental health episode? Perhaps add specific details to clarify.

REVIEWER	Reviewer name: Ian Wright Institution and Country: University of Wollongong Australia Competing interests: Published several similar reviews on NAPLAN and have further application in process
REVIEW RETURNED	10-Jul-2019

GENERAL COMMENTS	The intro needs a wider global context If there is an intention to imply causation then the cause MUST precede the outcome Similarly if there are known preexisting social, birth characteristics, early exposures that are known to predict your outcomes AND predict you diagnoses as well as NAPLAN, then you at least need to extract this data to be able to account for it in modelling. Using whole population rather than case control may allow this modelling to be done more efficiently

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Comments to the Author

I have performed a statistical review of the manuscript "Impact of chronic health conditions and injury on school performance and health outcomes in New South Wales, Australia: a record linkage study protocol" The authors describe a data linkage study with the objective of identifying the impact of injury and chronic illness on school performance. Overall the protocol is clear and well written. I have two comments on the analysis approach and a few comments on wording

Content comments

1) Record linkage: it is certainly valuable to have a third party conducted in data linkage. My comment relates to the final ‘linked’ dataset which will be provided to the authors for this work.

I assume that the identifying information (names, date of birth) used by the third party for data linkage will not be made available to the authors of this work for protection of participant privacy?

Response: Yes, that is correct; name, address or date of birth are not provided to the investigators and are only used by the third party for record linkage purposes. That identifying information is not provided to investigators is indicated in line 9 of the section entitled ‘Record linkage’.

Has any consideration also been given to potentially ‘sensitive’ data which may be present within the databases? Particularly relating to mental health? Will any additional steps be taken either by the third party responsible for linkage or the present authors to minimise any risks of data breach or re-identification of children with potentially sensitive information disclosed?

Response: Yes, to prevent any potential identification of individuals, investigators will not report information that includes small cell sizes – i.e. cell sizes less than 5. This information has been added to the 'Data analysis plan' section.

2) Data analysis plan: there are a few areas here I am unsure on, could the authors clarify?

a) Page 9, line 36: "Child injury and each chronic illness will be examined separately." So does this mean that for the primary and secondary outcomes, a separate regression model will be used for hospitalisations due to injury and due to each of the four chronic conditions?

Response: Yes, that is correct, as specified in the 'Data analysis plan' section, injury and each chronic illness will be examined separately.

b) Generalized linear regression is proposed for the primary outcome of 'proportion of performances below the NMS.' Linear regression methods require continuous data to be normally distributed, consider whether this 'proportion' would be normally distributed. Perhaps using a logistic regression for NAPLAN domain score below the NMS (yes or no) – I think this is the approach described over lines 43-49.

Response: Logistic regression is a type of generalized linear regression. There is a difference between general linear regression (GLM) and generalized linear regression (GZLM). For GLM, the response variable needs to be continuous, but for GZLM the response variable can also be categorical. Unlike GLM, GZLM does not require the assumption of normal distribution.

c) Related to the above comment, I suggest that hospital length of stay and hospital treatment costs are unlikely to be linear (page 10, line 25). Such data is generally very positively skewed and influenced by extreme values (longer and more costly hospital stays than expected. Consider methods which allow for the likely 'non-linearity' (i.e. skewed) nature of this information

Response: See response above, generalized linear regression does not require the assumption of normal distribution.

d) Page 9, line 51: Why will both relative risks and odds ratios reported? Output of logistic regression models would be odds ratios.

Response: The reviewer is correct. It is true that logistic regression does not usually produce relative risks directly, but it is possible to produce relative risks using PROC NLMIXED or by using the macro outlined in SAS documentation regarding estimating a relative risk - http://support.sas.com/kb/23/003.html

To produce both odds ratios and relative risk estimates more directly, the investigators can use log-binomial regression in PROC GENMOD, if the model converges.

e) Page 9, line 53: Please add further details on sensitivity analyses for potential missing data values (for example any imputation methods?)

Response: Additional information has been included in the 'Data analysis plan' section on sensitivity analyses for potential missing data values. Potential missing values will be imputed using the discriminant function method with 100 imputations using PROC MI. Parameter estimates will be log-transformed and pooled results and 95%CIs will be generated using PROC MIANALYSE. Analyses will be performed with and without imputed data. Imputing up to 30% missing data in a sample ≥5000 has been reported as acceptable (Meeyai, S., Logistic regression with missing data: A comparison of handling methods, and effects of percent missing values. Journal of Traffic and Logistics Engineering, 2016. 4(2): p. 128-134).

Wording comments

1) What this study adds “This study will identify the types of injuries and chronic illness associated with problems with learning at school” Unless I’ve missed something in this protocol, I didn’t think that the ‘type’ of injury was considered, just whether the child had been hospitalised due to an injury or not (and injury severity from Table 2). Please clarify and maybe reword.

Response: The authors will be able to consider different types of injuries within the analysis, depending upon sample size. Eg traumatic brain injury, burns, orthopaedic injury. Paragraph 2 of the ‘Introduction’ indicates that different types of injuries may affect a child’s health in different ways, particularly injuries that are serious. Additional information has been added into the ‘Data analysis plan’ section to indicate that some types of injuries, such as traumatic brain injury, may be examined separately, depending on sample size.

2) There are a few references throughout to ‘years’ and ‘grades’ of the children (e.g. page 5, objective 3; page 9, line 42). I presume these years / grades map specifically to the Australian school system? E.g. I know that the English school years are different. As the readership of the journal may be global, I suggest instead to refer to ages of children (or at least defining the ages of the children within the different school years and grades mentioned here) for clarity.

Response: The ages of the children have been defined in the section on ‘Scholastic performance.’ ie. the National Assessment Plan for Literacy and Numeracy assessments are conducted on all Australian children in primary school years 3 (7-9 years of age) and 5 (9-11 years of age), and secondary school years 7 (11-13 years of age) and 9 (13-15 years of age).

3) Page 5, line 33-34: “These five conditions were selected as…” This wording didn’t quite seem right to me as the focus is on hospitalisation due to injuries or four listed chronic conditions – injuries are not a condition so it isn’t really five conditions. Consider rewording

Response: This sentence has been reworded to: “ These four health conditions and injury were selected as injuries are the leading cause of hospitalisation in Australia for children aged 1-18 years..”

4) Page 7, line 49: “The comparison group will be randomly matched in a 1:4 ratio on age, gender and residential postcode to their matched case.”

How exactly are ‘residential postcodes’ defined here and are these areas quite broad? For example, I’m fairly confident that in the UK, it would be very restrictive to try and match age gender and exact residential postcodes – whereas if the first half of the postcode was used in the UK, this would result in a much broader sampling area.

Response: In Australia, the residential postcodes generally cover a fairly broad area. Metropolitan postcodes can include one large suburb or several smaller suburbs. In regional areas, one postcode could include around 20+ suburbs/towns. The chief investigator has used age, gender and postcode to conduct a case-comparison study previously in four Australian states and the matching was able to be conducted using these three criteria, except there was a small sample to select from for the older age groups and we ended up matching on an age group of 85+ years, rather than single units of age.

5) Page 9, line 34: “All hospital episodes of care related to the one event will be linked to form a period of health care.” I’m not completely following what ‘one event’ means here. Does this refer to all hospitalisations for the same injury, or related to one epileptic seizure, one asthma attack, one hypoglycaemic episode or one mental health episode? Perhaps add specific details to clarify.

Response: Yes, the reviewer is correct. New South Wales hospitalisation data is recorded as episodes of care (e.g. this injury event has 5 episodes of care: one record for admission to ICU, a second record for transfer to ward, then a third for transfer back to ICU, then a fourth for transfer to ward, and a fifth record for transfer to rehabilitation). All of these episodes of care for the one injury event need to be linked and analysed as one 'period of care'. An example has been added to the 'Data analysis plan' section – eg all episodes of care related to the same injury event.

Reviewer: 2

Comments to the Author

The intro needs a wider global context

Response: Additional information regarding the global context has been added to the 'Introduction' section regarding the importance of receiving good primary and secondary education for children and adolescents, as specified in the World Health Organization global strategy for child and adolescent health. References have also been added to indicate that injury is one of the leading causes of hospitalisation for children worldwide and that chronic health conditions are also prevalent among children globally.

If there is an intention to imply causation then the cause MUST precede the outcome

Response: This study will examine associations between injury and the health conditions and school performance. It will not be possible to examine causation.

Similarly if there are known preexisting social, birth characteristics, early exposures that are known to predict your outcomes AND predict you diagnoses as well as NAPLAN, then you at least need to extract this data to be able to account for it in modelling.

Response: The investigators agree and have tried to access as many potential mediating and explanatory data variables as possible that are recorded within the available administrative data collections that are accessible for record linkage in New South Wales. The investigators will be able to consider information pertinent to the child, their parents and clinical factors that all could impact on educational performance.

Using whole population rather than case control may allow this modelling to be done more efficiently

Response: At this stage, the cases are being identified from the total population of children who have been hospitalised in New South Wales and the comparison group is being selected and matched to the cases from all the children born in New South Wales. If the reviewer intended for all children who had not been hospitalised with an injury or one of the four health conditions to be included in the comparison group, the cost of the record linkage would have been prohibitive for the investigators. There are an estimated 1.7 million children aged ≤18 years in NSW whose data would need to be linked annually.

VERSION 2 – REVIEW

REVIEWER	Reviewer name: Sarah Nevitt Institution and Country: University of Liverpool Competing interests: I have no competing interests
REVIEW RETURNED	07-Aug-2019

GENERAL COMMENTS	Thank you to the authors for their responses and clarifications relating to my statistical comments. I am satisfied that all of my comments have been addressed and I am happy to recommend this manuscript for publication