## PEER REVIEW HISTORY



## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | Childhood respiratory illness presentation and service utilisation in primary care: a six year cohort study in Wellington, New Zealand using Natural Language Proccessing (NLP) software. |
| **AUTHORS** | Dowell, Anthony; Darlow, Ben; MacRae, Jayden; Stubbe, Maria; Turner, Nikki; McBain, Lynn |

## VERSION 1 - REVIEW

| | |
|---|---|
| **REVIEWER** | Kerry Hall<br>QUT, Australia |
| **REVIEW RETURNED** | 09-May-2017 |

| | |
|---|---|
| **GENERAL COMMENTS** | Thank you for asking me to review this paper on childhood respiratory illness and service utilisation in primary care. It Respiratory illness are and well known cause of childhood mortality globally, therefor an understanding of health service utilization for respiratory illness at the primary care level, could be an important factor in improving health care delivery at the primary care level. I do have some concerns about the paper firstly, it is a retrospective cohort study, it does not adhere to the STROBE guidelines for the structure of a cohort study. Also there is no evidence of ethics committee approval, I have some concerns reviewing this paper unless an ethics committee approval can be provided. |

| | |
|---|---|
| **REVIEWER** | Ross, Mindy<br>UCLA |
| **REVIEW RETURNED** | 13-May-2017 |

| | |
|---|---|
| **GENERAL COMMENTS** | Thank you for the opportunity to review the manuscript bmjopen-2017-017146. This project applies natural language processing to clinical progress notes to document health care utilization for respiratory tract illness in the pediatric primary care setting.<br><br>General<br>1) Overall, the descriptive statistics about respiratory illness confirm knowledge that exists but is still publishable in that this data has not been documented in detail from EHR records in New Zealand (or potentially elsewhere). However, this manuscript recycles the group's previously published manuscript to too great of an extent to seem acceptable for publication in its current stat (MacRae, J., et al. "Accessing primary care Big Data: the development of a software algorithm to explore the rich content of consultation records." BMJ open 5.8 (2015): e008160). The title is too similar to the previous publication so if this is accepted, consider changing the title of the paper to mention natural language processing (NLP) instead of "big |

| | |
|---|---|
| | data." Using the term "big data" is somewhat vague. In addition, including the term NLP will be more helpful to the reader or anyone who comes across the paper. A key word could be "big data" instead.<br>2) Unless you make it clear that "otitis media" is a complication of a viral infection, you should remove "otitis media" because it's not a respiratory illness (SNOMED-CT parent is "disorder of middle ear"). In addition, you may want to say "respiratory tract" illness because pharyngitis does not affect gas exchange but is a part of the upper airway (respiratory tract).<br>3) The abstract says you use free text, but in the methods, it appears that diagnosis codes and prescribing information are also used (discrete data)? Perhaps that can be clarified more.<br>Also, I'm not familiar with "Read codes," but I think they're the same as ICD (International Classification of Disesases) codes? Do you know how much more accurate at identifying cases is this method is versus trying it with the Read/ICD codes alone?<br>4) The introduction may need more research work/references. In the introduction, are you referring to lack of primary care utilization data in general and for respiratory illness in New Zealand or throughout the world? If it's New Zealand, please clarify. If it's worldwide, there are many studies. For example, a PubMed search of: "Respiratory tract diseases/epidemiology"[Majr]) AND "Primary Health Care"[Majr] returns >300 papers and "Respiratory tract diseases/epidemiology"[MeSH]) AND "Primary Health Care"[MeSH] returns >1400 papers.<br>5) Minor- make sure to spell out all abbreviations before they're abbreviated (e.g. GP and OECD in the abstract). Also, if abbreviations are allowed in the abstract, once you use a term such as "electronic medical record" then can abbreviate EMR.<br><br>Specific Comments<br>Abstract/Strength-Limitations<br>• Page 2, line 9: Re: "respiratory illness" -- See #2 above. I will only reference this once but it applies throughout.<br>• Page 2, line 13: Clarification about the term "enrolled." It seems that clinics were enrolled, but not individual patients. The way it is worded here, it reads as if the individual patients were enrolled, which I don't think is the case. Was this an Institutional Review Board (IRB) waiver of consent study?<br>• Page 2, lines 33-39: Re: "otitis media" and "throat infection," see #2 above.<br>• Page 2, lines 43-49: Have there been large studies using EHR Read/ICD codes alone?<br>• Page 3, lines 5-6: To be more specific for the reader, instead of (or after the term) "big data," it would be more clear to add a line like on page 21, line 22-23 about the NLP algorithm exploring structured and unstructured data from consultation notes<br>• Page 3, line 10: Please clarify what's meant by "high degree of accuracy" as the next sentence and weaknesses of the study seems to contradict a bold statement like this.<br>• Page 3, line 15: I'd break up this bullet point. The NLP methodology is different than application to other diagnoses. Again, using the term "big data" is vague.<br><br>Introduction<br>• Too similar to previous paper mentioned in point #1 of "General" section<br>• Page 4, line 6. Consider reference such as Bethell, Christina D., et al. "Adverse childhood experiences: assessing the impact on health |

| | |
|---|---|
| | and school engagement and the mitigating role of resilience." Health Affairs 33.12 (2014): 2106-2115.<br>• Page 4, line 8-9. This sentence seems a bit overreaching. There is data from each institution, but maybe not published. Also, this sentence is contradicted somewhat by sentence on this page starting on lines 21-22 and reference 16.<br>• Page 4, lines 10-12. There have to be more than a few papers on "respiratory illness" if it includes all conditions in the table on Appendix 1. I would remove this sentence and start with "Children under five…"<br>• Page 4, line 13: Same comment in regard to using "otitis media" #2 in general comments<br>• Page 4, lines 42-43: Reference(s) after sentence that ends with …available (primarily clinical consultation notes). There are other NLP tools in existence that can analyze clinical notes. The follow-up sentence seems to say that New Zealand for some reason was unable to use any NLP tools prior to the [ref 17 and 18], but I don't think that's true, is it?<br><br>Methods<br>• Too similar to previous paper mentioned in point #1 of "General" section<br>• Page 4. Line 53. Would consider starting this section with "Setting and Participants," then follow with Design and Process<br>• Page 4, Settings and Participants. After reading the [ref 17] paper, it appears that this section was done in the previous study? You write this section as if this was done just for this paper. But, the numbers in this manuscript are slightly different than [ref 17], so it's unclear to the reader how much of this was done before or if this was repeated for this paper.<br>• Page 4, Process Section (lines 32-60). Again, it appears that this section copies [ref 17]<br>• Page 5, line 41-42. This last sentence isn't clear…if they have a respiratory illness finding, like "wheeze," on physical exam but they came into the clinic for "gastroenteritis" then they would be excluded?<br>• Page 6, lines 3-6. Appendix 2 is the exact same as Table 3 from [ref 17]<br><br>Results<br>• Appendix 3 is exactly the same as Figure 3 in [ref 17]<br>• For text and figures 2-5 -- See general comment #2 re: otitis media and throat infection<br>• Page 7, lines 9-14. When you reference figure 6, the clinical reader might be interested in a summary of the average number of respiratory infections per year the patients in the first 2 years of life have compared to older children and adolescents. They can look at the figure, but might be nice to mention here briefly. This is to compare it to some references that state "normal" number of respiratory illness for young kids could be 6-8 per winter season.<br><br>Discussion<br>• Page 10, line 37-40. Worldwide or in New Zealand? May be the first study this large directly from EHR data?<br>• Page 11, line 33-34. The end of this sentence seems to be contradicted by the limitations to the method. |

**VERSION 1 – AUTHOR RESPONSE**

Reviewer: 1
I do have some concerns about the paper firstly, it is a retrospective cohort study, it does not adhere to the STROBE guidelines for the structure of a cohort study.
We have gone through the check list to ensure this study covers all requirements – refer table above

Also there is no evidence of ethics committee approval, I have some concerns reviewing this paper unless an ethics committee approval can be provided.
Ethics approval attached

Reviewer Two

General
1) Overall, the descriptive statistics about respiratory illness confirm knowledge that exists but is still publishable in that this data has not been documented in detail from EHR records in New Zealand (or potentially elsewhere). However, this manuscript recycles the group's previously published manuscript to too great of an extent to seem acceptable for publication in its current stat (MacRae, J., et al. "Accessing primary care Big Data: the development of a software algorithm to explore the rich content of consultation records." BMJ open 5.8 (2015): e008160). The title is too similar to the previous publication so if this is accepted, consider changing the title of the paper to mention natural language processing (NLP) instead of "big data." Using the term "big data" is somewhat vague. In addition, including the term NLP will be more helpful to the reader or anyone who comes across the paper. A key word could be "big data" instead.
We have reviewed our text where there is similar language, acknowledged further the earlier publication around design methodology and altered the title to NLP thank you.
P1 Title
P 5 – setting and participants.
P 11 – strengths and limitations.

2) Unless you make it clear that "otitis media" is a complication of a viral infection, you should remove "otitis media" because it's not a respiratory illness (SNOMED-CT parent is "disorder of middle ear"). In addition, you may want to say "respiratory tract" illness because pharyngitis does not affect gas exchange but is a part of the upper airway (respiratory tract).
We have added in the word 'tract' to respiratory illness and have altered the wording through the manuscript to recognise that otitis media is not a direct respiratory illness itself, but a complication related to respiratory tract illness.
P 5 – Process relates to description of otitis media

3) The abstract says you use free text, but in the methods, it appears that diagnosis codes and prescribing information are also used (discrete data)? Perhaps that can be clarified more.
Also, I'm not familiar with "Read codes," but I think they're the same as ICD (International Classification of Disesases) codes? Do you know how much more accurate at identifying cases is this method is versus trying it with the Read/ICD codes alone?

In the abstract we state that we also use coded data.
P 2 Line 11,12
In the methods section we also make it clear that the algorithm uses combined sources of information.
P5 Line 42/43

Read codes are the standard coding system used within primary care in many countries, and were developed from the Primary care version of ICD. We had not described or referenced the Read coding system, as it appears unreferenced in many journals including BMJ Open.
A suitable reference if required would be:
Chisholm J. The Read clinical classification. BMJ: British Medical Journal. 1990 Apr 28;300(6732):1092.

We have referenced previous research which has found advantages of natural language processing accuracy over diagnostic codes, in the strengths and limitations section of the paper.
Wu ST, Sohn S, Ravikumar KE, et al. Automated chart review for asthma cohort identification using natural language processing: an exploratory study. Ann Allergy Asthma Immunol 2013; 111(5): 364-9.
P11 Line 42/43

3) The introduction may need more research work/references. In the introduction, are you referring to lack of primary care utilization data in general and for respiratory illness in New Zealand or throughout the world? If it's New Zealand, please clarify. If it's worldwide, there are many studies. For example, a PubMed search of: "Respiratory tract diseases/epidemiology"[Majr]) AND "Primary Health Care"[Majr] returns >300 papers and "Respiratory tract diseases/epidemiology"[MeSH]) AND "Primary Health Care"[MeSH] returns >1400 papers.

Thank you.
We believe the reviewer has been focusing on the broader topic of respiratory tract disease in general in primary care; however the relevant literature here is related to describing the burden of respiratory tract illness in primary care where there is much more limited published data.
While PubMed searching reveals a large number of putative references, combination of respiratory tract epidemiology and primary health care produced a list of 31, of which the majority were of small individual practices, used disease codes to assess prevalence or reason for encounter, or were not related to paediatric practice. None of the references used Natural Language Processing methodology except a reference to our previous methodology paper which did not address the clinical or service utility aspects.

We have changed a sentence in the introduction to reflect the availability of general epidemiological studies. "Respiratory illness contributes substantially to childhood morbidity yet despite the plethora of studies of general respiratory epidemiology few data exist describing the burden of respiratory tract related illness in routine primary care."
P 4 Line 10-12

In the discussion we have added additional references and text to reflect the nature of some of the other studies identified.
P 10 Line 53
Epidemiología de las consultas pediátricas respiratorias en Santiago de Chile desde 1993 a 2009. Revista Panamericana de Salud Pública 32(1): 56-61.

P10 Line 56
Rosa, A. M., et al. (2008). "Respiratory disease and climatic seasonality in children under 15 years old in a town in the Brazilian Amazon." Jornal de Pediatria 84(6): 543-549.

P11 Line 10/11
Molony, D., et al. (2016). "70,489 primary care encounters: retrospective analysis of morbidity at a primary care centre in Ireland." Irish Journal of Medical Science (1971-) 185(4): 805-811.

4) Minor- make sure to spell out all abbreviations before they're abbreviated (e.g. GP and OECD in the abstract). Also, if abbreviations are allowed in the abstract, once you use a term such as "electronic medical record" then can abbreviate EMR.
Have checked and altered where needed

Specific Comments
Abstract/Strength-Limitations
• Page 2, line 9: Re: "respiratory illness" -- See #2 above. I will only reference this once but it applies throughout.
As above, we have altered throughout

• Page 2, line 13: Clarification about the term "enrolled." It seems that clinics were enrolled, but not individual patients. The way it is worded here, it reads as if the individual patients were enrolled, which I don't think is the case. Was this an Institutional Review Board (IRB) waiver of consent study?
Individuals are enrolled with clinics, and the clinics are enrolled with the study. As there was no identifiable individual data the Ethics Committee approved the methodology.

• Page 2, lines 33-39: Re: "otitis media" and "throat infection," see #2 above.
As per previous comments, we have altered the language in this manuscript in line with these concerns

• Page 2, lines 43-49: Have there been large studies using EHR Read/ICD codes alone?
This study is focused on incidence and service utilisation. All known published studies relevant to this are included in our introduction

• Page 3, lines 5-6: To be more specific for the reader, instead of (or after the term) "big data," it would be more clear to add a line like on page 21, line 22-23 about the NLP algorithm exploring structured and unstructured data from consultation notes
As NLP is mentioned in the first half of this sentence and it is fully described in the introduction, we have chosen to leave the word 'big data' in here.

• Page 3, line 10: Please clarify what's meant by "high degree of accuracy" as the next sentence and weaknesses of the study seems to contradict a bold statement like this.
We consider that a 'high degree of accuracy' around assessing the burden of disease is not in conflict with generating a 'conservative estimate'; and feel the paper is clear about that in our results and discussion

• Page 3, line 15: I'd break up this bullet point. The NLP methodology is different than application to other diagnoses. Again, using the term "big data" is vague.
We have altered the wording here to remove 'big data'.

Introduction
• Too similar to previous paper mentioned in point #1 of "General" section
As above, we have reviewed and altered the overlap

• Page 4, line 6. Consider reference such as Bethell, Christina D., et al. "Adverse childhood experiences: assessing the impact on health and school engagement and the mitigating role of resilience." Health Affairs 33.12 (2014): 2106-2115.

Thank you. We have added this reference.

• Page 4, line 8-9. This sentence seems a bit overreaching. There is data from each institution, but maybe not published. Also, this sentence is contradicted somewhat by sentence on this page starting on lines 21-22 and reference 16.
We have not done a review of non-published ('grey') data and we cannot be sure of robustness of results without publication, so we have added in a ref to the fact we are referring to 'published' data. Lines 21- 22 do refer to international data and yet that does not invalidate the fact there is limited data.

• Page 4, lines 10-12. There have to be more than a few papers on "respiratory illness" if it includes all conditions in the table on Appendix 1. I would remove this sentence and start with "Children under five…"
As per our earlier comments this is in reference to the burden of respiratory illness in childhood in primary care.

• Page 4, line 13: Same comment in regard to using "otitis media" #2 in general comments
We acknowledge as per our earlier comments thank you

• Page 4, lines 42-43: Reference(s) after sentence that ends with …available (primarily clinical consultation notes). There are other NLP tools in existence that can analyze clinical notes. The follow-up sentence seems to say that New Zealand for some reason was unable to use any NLP tools prior to the [ref 17 and 18], but I don't think that's true, is it?
The sentence does state in relation to New Zealand that there has been ."some exploration previously but not to analyse childhood respiratory service utilisation due to difficulties with extracting and analysing both structured and unstructured data available …' This is correct.

Methods
• Too similar to previous paper mentioned in point #1 of "General" section
Thank you, altered as per our earlier comments

• Page 4. Line 53. Would consider starting this section with "Setting and Participants," then follow with Design and Process
Thank you, we have considered this but still feel the initial sentence explaining the design is useful prior to explaining the setting and participants

• Page 4, Settings and Participants. After reading the [ref 17] paper, it appears that this section was done in the previous study? You write this section as if this was done just for this paper. But, the numbers in this manuscript are slightly different than [ref 17], so it's unclear to the reader how much of this was done before or if this was repeated for this paper.
The methodology was as described in the original paper, but it is correct that this is a new study based on the design, hence the numbers are not the same. We hope the improved writing with less overlap has helped to clarify this

• Page 4, Process Section (lines 32-60). Again, it appears that this section copies [ref 17]
As per our earlier comments re reviewing overlap

• Page 5, line 41-42. This last sentence isn't clear…if they have a respiratory illness finding, like "wheeze," on physical exam but they came into the clinic for "gastroenteritis" then they would be excluded?

Thank you. We have changed the wording of sentences in the process section to make clear that 'respiratory included all respiratory tract-related conditions and presentations' and 'non respiratory included consultations where the primary presentation and diagnosis was for conditions such as injury

or gastroenteritis'.

• Page 6, lines 3-6. Appendix 2 is the exact same as Table 3 from [ref 17]

Thank you. We felt that in this case it would be helpful for readers to have an idea of the performance of the algorithm during its validation. We are aware of the Editors comments about the use of text from our previous publication, and have thus removed .

Results
• Appendix 3 is exactly the same as Figure 3 in [ref 17]

Table 3 is not the same as Fig 3 in reference . Appendix 3 refers to the Demographic characteristics of children in the study cohort at the end of the study compared with enrolled population and national census data for this study. Figure 3 from the previous publication refers to the Demographic characteristics the practice populations that were included in the validation of the NLP algorithm.

• For text and figures 2-5 -- See general comment #2 re: otitis media and throat infection
Thank you ,we have addressed these

• Page 7, lines 9-14. When you reference figure 6, the clinical reader might be interested in a summary of the average number of respiratory infections per year the patients in the first 2 years of life have compared to older children and adolescents. They can look at the figure, but might be nice to mention here briefly. This is to compare it to some references that state "normal" number of respiratory illness for young kids could be 6-8 per winter season.

We have provided additional information summarising the number of respiratory infections in different ages of the cohort.
The mean number of presentations for respiratory tract infection for an individual was 2.6 per year in those under 2 years, 2.1 per year in those aged 3 to 5 years, and 1.5 per year in those over 15 years.

Discussion
• Page 10, line 37-40. Worldwide or in New Zealand? May be the first study this large directly from EHR data?
Based on our literature review we feel that this sentence is correct when it states that 'this is the first study to assess the primary care incidence and service utliisation….in such a large cohort…" and we cannot find any published literature to disprove this so feel it is reasonable to leave the sentence as it is written.

• Page 11, line 33-34. The end of this sentence seems to be contradicted by the limitations to the method.
As our comments above we consider that a 'high degree of accuracy' around assessing the burden of disease is not in conflict with generating a 'conservative estimate';