BMJ Open

Childhood respiratory illness presentation and service utilisation in primary care: a six year cohort study using big data.

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-017146
Article Type:	Research
Date Submitted by the Author:	04-Apr-2017
Complete List of Authors:	Dowell, Anthony; University of Otago - Wellington , Primary Health Care and General Practice Darlow, Ben; University of Otago, Primary Health Care and General Practice MacRae, Jayden; Datacraft Analytics Stubbe, Maria; University of Otago, Primary Health Care and General Practice Turner, Nikki; University of Auckland, General Practice and Primary Health Care McBain, Lynn; University of Otago, Wellington,
Primary Subject Heading:	General practice / Family practice
Secondary Subject Heading:	Health informatics, Paediatrics, Respiratory medicine
Keywords:	PRIMARY CARE, General Practice, Child health, Big Data, Respiratory illness

Childhood respiratory illness presentation and service utilisation in primary care: a six year cohort study using big data.

Anthony Dowell. MBChB. – corresponding author.
Department of Primary Health Care and General Practice.
University of Otago – Wellington.
23 Mein St, Newtown, Wellington 6242. New Zealand.
Tony.dowell@otago.ac.nz
+64 21 270 1617.

Ben Darlow. PhD
Department of Primary Health Care and General Practice.
University of Otago – Wellington. 23 Mein St, Newtown, Wellington 6242. New Zealand.

Jayden Macrae. MsC
Patients First, Level 4, 50 Customhouse Quay, Wellington 6011, New Zealand.

Maria Stubbe. PhD.
Department of Primary Health Care and General Practice.
University of Otago – Wellington. 23 Mein St, Newtown, Wellington 6242. New Zealand.

Nikki Turner. MD.
Department of General Practice and Primary Health Care. University of Auckland. Level 8, Petherick Towers. 38 Waring Taylor St. Wellington CBD. New Zealand.

Lynn McBain. MD.
Department of Primary Health Care and General Practice.
University of Otago – Wellington. 23 Mein St, Newtown, Wellington 6242. New Zealand.

Word count 2955

Key words
Primary Care, General Practice, Childhood respiratory illness, Natural language software programming, Big data.

Abstract

Objectives.

To identify childhood respiratory illness presentation rates and service utilisation in primary care, by interrogating free text and coded data from Electronic Medical Records.

Design.

Retrospective cohort study. Data interrogation used a natural language processing software inference algorithm.

Setting. 36 primary care practices in New Zealand. Data analysed from January 2008 – December 2013.

Partcipants

The records from 77,582 children enrolled in were reviewed over a six-year period to estimate the presentation of childhood respiratory illness and service utilisation.This cohort represents 268,919 person years of data and over 650,000 unique consultations.

Main outcome measure. Childhood respiratory illness presentation rate to Primary Care practice, with description of seasonal and yearly variation.

Results.

Respiratory conditions constituted 46 per cent of all child-GP consultations with a stable year on year pattern of seasonal peaks. Upper Respiratory Tract Infection was the most common respiratory category accounting for 21.0% of all childhood consultations, followed by otitis media (12.2%), wheeze-related illness (9.7%), throat infection (7.4 %), and Lower Respiratory Tract Infection (4.4 %). Almost 70 per cent of children presented to their GP with at least one respiratory condition in their first year of life; this reduced to approximately 25 per cent for children aged 10 to 17.

Conclusion.

This is the first study to assess the primary care incidence and service utilisation of childhood respiratory illness in a large primary care cohort by interrrogating Electronic Medical Record free text. The study identified the very high primary care workload related to childhood respiratory illness, especially during the first two years of life. These data can enable more effective planning of health service delivery. The findings and methodology have relevance to many OECD countries, and the use of primary care 'big data' in this way can be applied to other health conditions.

Strengths and limitations of this study

- This study uses a novel and validated natural language processing software inference algorithm to identify childhood respiratory illness presentation rates and service utilisation using primary care Big Data.
- The presentation and burden of childhood respiratory diseases in primary care has not previously been estimated with such a high degree of accuracy.
- The algorithm was designed to maximise specificity, thereby generating a conservative estimate of the burden of childhood respiratory disease in primary care by keeping false positives to a minimum
- The methodology has relevance to many OECD countries, and the use of primary care 'big data' in this way can be applied to other health conditions.
- This study analysed normal hours primary care GP consultations. The exclusion of nurse-only and out-of-hours consultations may result in an underestimation of primary care respiratory presentation rates.

Introduction

Childhood is a crucial period for development and well-being. A healthy start to life reduces adulthood morbidity and enhances participation in society.[1-4] Physical illness is an important risk factor for poor health outcomes.[5] Globally, primary care is utilised by all children,[6] but there is currently little knowledge of detailed morbidity and utilisation patterns in community settings.

Respiratory illness contributes substantially to childhood morbidity yet few data exist describing the burden of respiratory illness in primary care. Children under five present up to six times a year with acute respiratory infections[7] and high prevalence rates are noted for asthma[8] and otitis media.[9] Such data are, however, mainly reliant on survey responses and parental report. These reports also lack precision regarding individual respiratory conditions, symptom severity, longitudinal patterns and variance related to age and seasonality. These data are needed to effectively plan primary health care service delivery. More detailed hospitalisation data are available[10]; however, these represent an unknown proportion of all cases and are based on diagnostic coding of uncertain accuracy.

International data suggest that respiratory conditions constitute 20 to 25% of all general practitioner (GP) consultations, with higher rates in those under 25 years.[11,12] These data are based on GP self-report, and accuracy may be limited by the competing demands of reporting, meeting patient needs and practice management tasks. Wide variance has been reported in how GPs describe the reason for encounter.[12]

Improved understanding of primary care childhood respiratory illness presentation could enable more systematic approaches to care and resource allocation, and a context for exploring important social and ethnic variations in hospitalisation rates.[10,13] In OECD countries, conditions such as bronchiolitis, asthma, upper respiratory tract infections, and pneumonia make up over 40 % of Ambulatory Sensitive Hospitalisation (ASH); admissions considered preventable through interventions delivered outside of hospitals, predominantly within primary care.[5,10,14,15]

More accurate assessment of illness presentation and service utilisation could be obtained by analysing consultation notes within electronic medical records (EMR) common in OECD primary care settings. While there has been some exploration of the potential for 'big data' assessment of general practice workload,[16] these data have not previously been used to analyse childhood respiratory service utilisation due to difficulties with extracting and analysing both structured and unstructured data available (primarily clinical consultation notes). The development of novel software has enabled the exploration of New Zealand EMR data.[17,18]

This study aimed to interrogate data from EMR to identify primary care presentation and service utilisation related to common childhood respiratory conditions.

Methods

Design

A natural language processing software inference algorithm was developed to interrogate quantitative and qualitative cross-sectional and retrospective cohort data from EMR.[17,18]

Setting and participants

Figure 1 illustrates the creation and analysis of the dataset. In New Zealand there is universal enrolment with a primary care practice. All 60 practices within the networks of two primary health organisations (PHOs) in the Wellington region of New Zealand were invited to participate and 36 consented. The study area contains mixed city, urban and rural settings and the cohort consisted of the 77,582 children (75% of the two PHOs' child population; N=103,333) under 18 years of age enrolled in these 36 primary care practices between 1 January 2008 until 31 December 2013. This cohort represented 268,919 person years; children both joined and left this cohort during the six-year study period (e.g. births, deaths, turning 18 years, or moving into or out of a consenting practice).

Data were collected directly from EMR using software which automates the extraction, and secure transmission of large data sets. The dataset comprised records from consultations generated during both standard office hours and out-of-hours practice. Data were extracted from the EMR for all child-GP consultations at consenting practices during the study period (n= 687,136). Each consultation record was identified using an individual's National Health Index (NHI) number. The NHI is a unique identifier assigned to every person who uses health services in New Zealand and enabled records to be matched between datasets. Consultations for which there were poor quality data (2439 consults from 256 children) were excluded. Out-of-hours consultations (n=34,584) were not analysed due to differing participation in out-of-hours services by the practices. All data were analysed within the PHO which has rigorous protocols in place to ensure patient confidentiality. No identifiable data were ever accessed by the research team.

Process

Each of the 650,123 clinical consultation notes was interrogated by a software inference algorithm and hierarchical classification system described previously.[18] The algorithm classified consultation records using: clinical information recorded by GPs and practice nurses, any recorded Read code diagnostic classifications, and prescribing information. The first level of the hierarchy divided all consultations into either 'respiratory' or 'not respiratory'. The 'not respiratory' category included consultations for presentations such as injury or gastroenteritis, and consultations in which the respiratory system was examined and screened, but no signs, symptoms, or diagnoses were recorded. These screening consultations were excluded so that the burden of respiratory illness estimate was not inflated by consultations which did not result from a respiratory illness.

The second level of the hierarchy sub-classified consultations into one or more specific respiratory categories. These categories were determined by a group of clinical experts; consideration was given to the degree to which conditions could be mapped to high prevalence (that which is common) and/or responsible for significant morbidity and hospitalisation (that which is important).The six categories were i) upper respiratory tract infections (URTI); ii) lower respiratory tract infections (LRTI); iii) wheeze-related illnesses; iv) throat infections; v) otitis media; and vi) other respiratory conditions. The conditions included within each diagnostic category are presented in Appendix 1.

The algorithm was trained, tested, and validated using three independent gold standard data sets of 1200 consultation records which had been independently classified by two general practice clinical experts (AD and LM). The algorithm was designed to replicate the judgements made by these clinical experts. Development aimed to optimise specificity while maximising sensitivity to minimise the

occurrence of false positives. The algorithm's sensitivity, specificity, positive and negative predictive values, and F-measure for each of the diagnostic categories against a gold standard validation set of 1200 consultation records are presented in Appendix 2.

Analysis

The demographic characteristics of age, gender, ethnicity (NZ indigenous Māori, Pacific, other), and New Zealand Deprivation Index (a measure of socioeconomic deprivation[19]) of the cohort (n=77,326) were compared with those of all children enrolled within the two PHOs (N=103,333) and the New Zealand population using national census data.

The proportion of primary care consultations for children aged 18 years and under which were related to the six specific respiratory conditions outlined above was obtained from the dataset using the algorithm. The utilisation of services for these six conditions was analysed by demographic characteristics. Consultation rates are expressed per 1000 child years observed due to the differing length of time individuals might be participants in the cohort. Patients were observed for the period in which they were enrolled in a participating practice; this was calculated from the date of a child's first visit to a practice until they were removed from the enrolment register. Both deprivation and ethnicity status were taken as the last ethnicity and deprivation recorded from the GP records. Consultation rates were adjusted for sensitivity and specificity of the algorithm (see Appendix 2) and a direct standardization method was applied to level 2 ethnicity and socio-economic deprivation quintiles against NZ Census 2013 data. Estimates of true rates were made using final test sensitivity and specificity results for each classification category using the method described by Rogan and Gladen. [20]All Data aggregation, transformation, cleaning and storage was done in Microsoft SQL Server, and statistical analysis was undertaken in R using packages including boot, epiR, combinat, stats, tm, RWeka, slam, SnowballC and caret. [21]

Results

The demographic characteristics of the study cohort closely matched those of the enrolled population (Appendix 3). The age distribution of the study cohort also closely matched the national comparison data. Compared with national census data the study cohort had a greater proportion of children from the least deprived quintile grouping (32% vs 25%) and a lower proportion of Māori (17% vs 22%).

From the 650,123 consultations reviewed the true rate of presentation for a respiratory condition was calculated to be 45.4 per cent of all consultations for children under 18 years of age (Figure 1). URTI was the most common respiratory category represented in 21.0% of all consultations, followed by otitis media (12.2%), wheeze-related illness (9.7%), throat infection (7.4%), and LRTI (4.4 %). Other respiratory classifications accounted for just 1.5% of all consultations . One respiratory condition was classified in 27.6% of all consultations, two in 7.0%, three in 0.8%, and greater than three in 0.1%. The rates of child respiratory condition consultation were 1,101 per 1,000 person years observed for all respiratory conditions, 509 per 1000 for URTI, 107 per 1000 for LRTI, 235 per 1000 for Wheeze Illness, 180 per 1000 for Throat Infections, 296 per 1000 for Otitis Media and 36 per 1000 for Other Respiratory conditions.The incidence of both respiratory and non-respiratory consultations remained stable throughout the study period with a consistent pattern of seasonal peaks and troughs (Fig 2). The respiratory consultation rate was highest in the Southern Hemisphere winter month of August, and lowest in January (Figure 3). Non-respiratory consultations followed a similar pattern but with shallower peaks and troughs. Respiratory conditions explained 64.4% of the

annual seasonal variation in child consultation rates. All respiratory conditions which were sub-classified followed a similar pattern of peaks in August and troughs in January except for 'other' respiratory conditions which were highest in December and lowest in April (Figure 4). Figures 4 and 5 present the annual variation in respiratory condition presentation for each classification category.

Respiratory consultations occurred throughout childhood, but at much greater frequency during the first two years of life. (Figure 6) During the first year of life 73.5 % of children presented to their GP with at least one respiratory condition. Following the second year of life, the presentation of all respiratory conditions decreased with increasing age. Of children aged 10 to 17 years, 22.5 % presented with at least one respiratory condition (Figure 6).

Fig 2 : Respiratory and non-respiratory consultations per quarter per 1000 enrolled children January 2008 to December 2013.

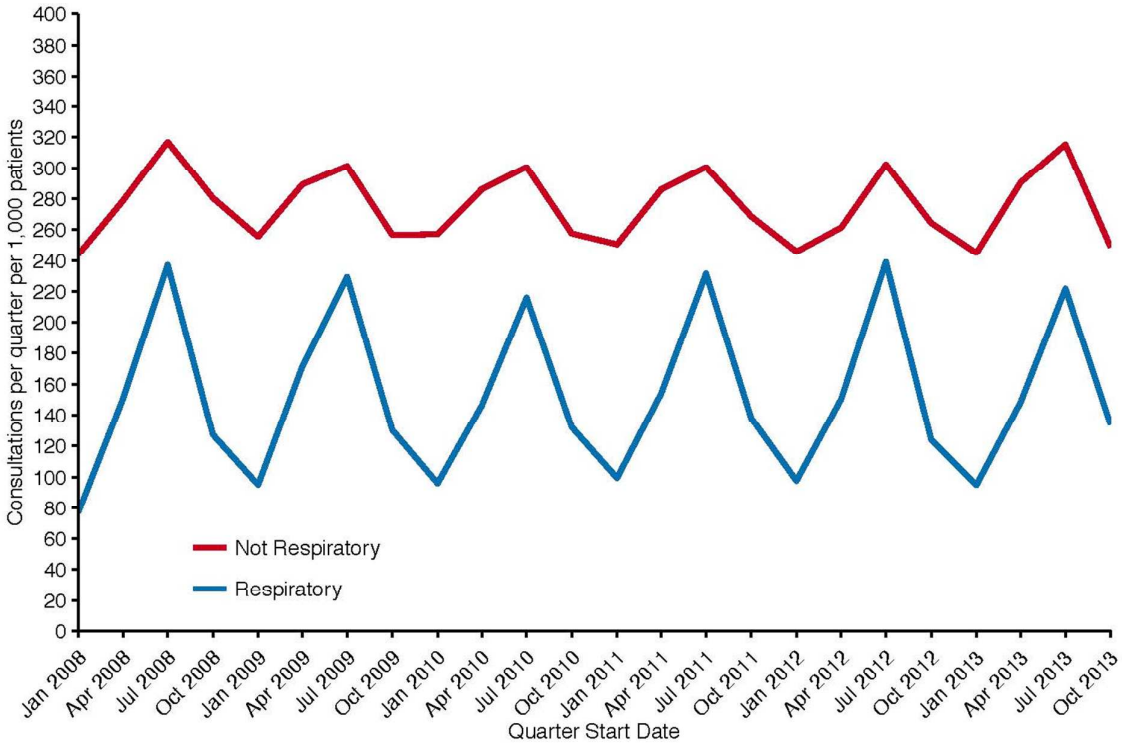


Figure 3: Mean respiratory and non-respiratory consultations per month per 1000 enrolled children. January 2008 to December 2013 demonstrating seasonal variation.

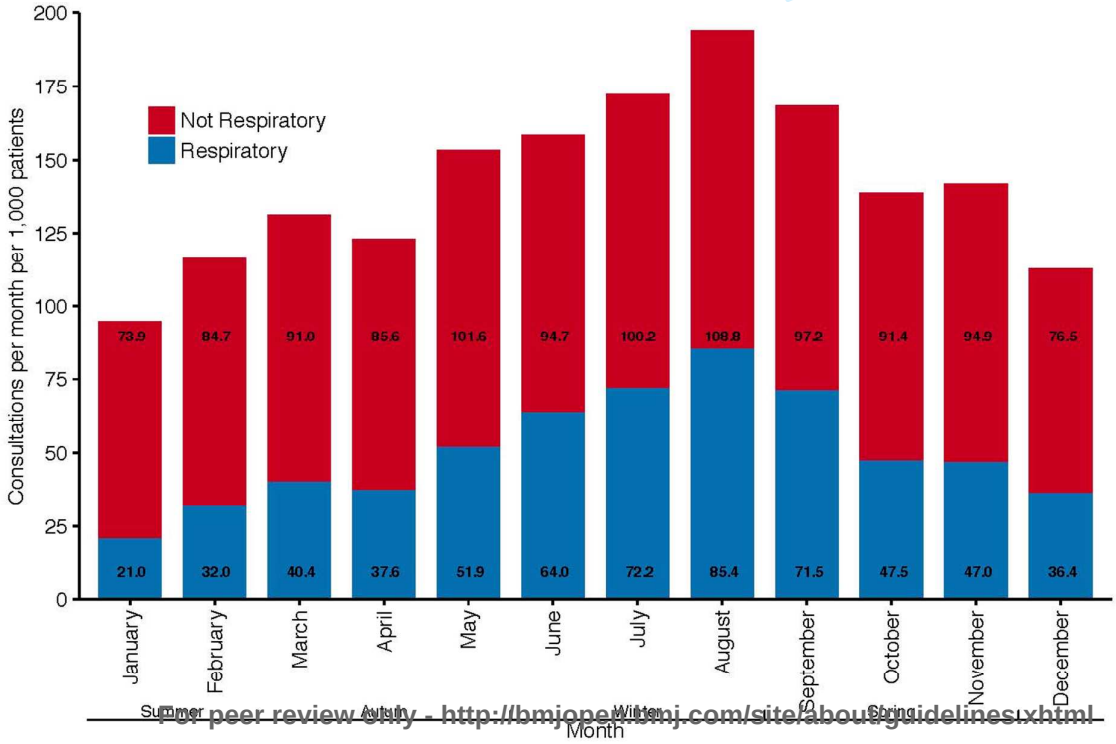


Figure 4: Yearly variation of consultations per month per 1000 enrolled children for each respiratory illness category - January 2008 to December 2013.

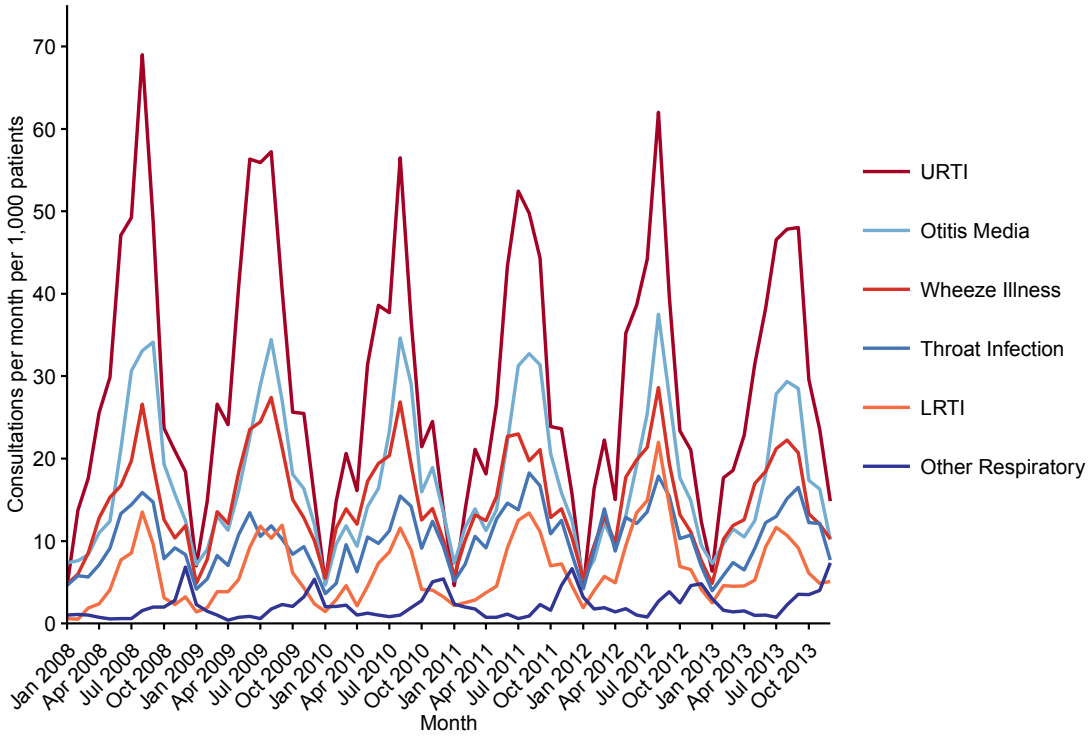


Figure 5 : Mean consultations per month per 1000 enrolled children for each respiratory illness category - January 2008 to December 2013

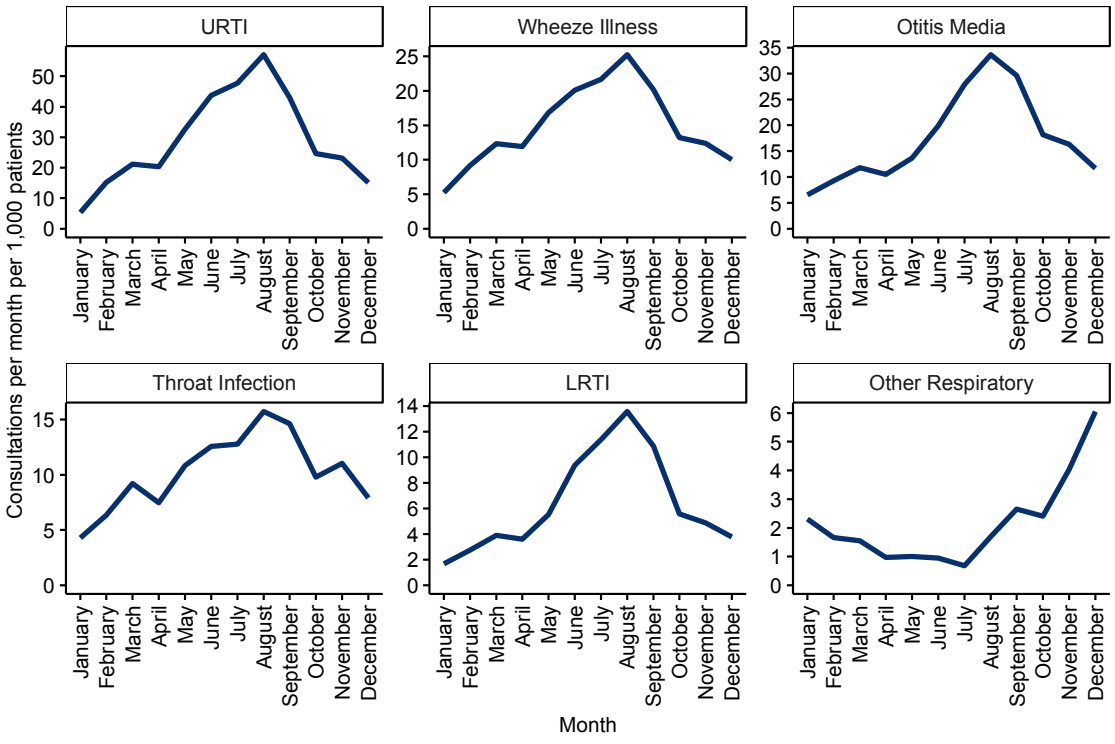


Figure 6: Respiratory consultation frequency by selected age cohort.

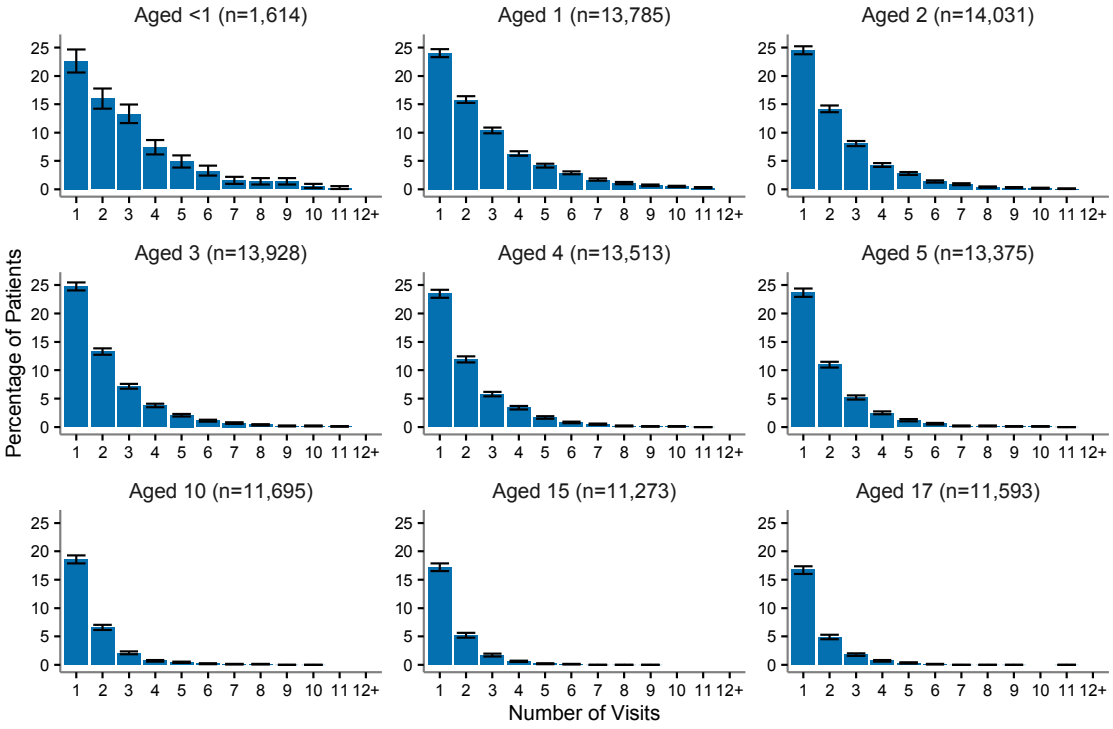


Figure 6: Each facet includes children who were enrolled within that age band for a twelve month period (e.g. from the day they turned one until the day before they turned two). The cohort of children under one is small because many children do not enrol until they are over three months old and may therefore only be enrolled for nine months before turning one.

Discussion

This is the first study to assess the primary care incidence and service utilisation of childhood respiratory illness in such a large cohort observing over 250,000 person years and more than 650,000 unique consultations. Using a novel and validated method of interrogating EMR free text, this study found that respiratory conditions constituted 45.4 per cent of all child-GP consultations. This quantifies the very high volume of childhood respiratory consultations and workload in general practice, especially during the first two years of life. These data can enable more effective planning of primary care service delivery and indicate areas in which to focus preventive programmes. The study also highlights the high presentation rates to primary care of those respiratory conditions which frequently present for hospital admission.

Comparison with other studies

The presentation rate of respiratory illness and pattern of seasonal peaks was remarkably stable across the six years included in this data set and was unchanged by events such as the H1N1 influenza pandemic of 2009. Consistent with findings from an Australian survey, the presentation of nearly all respiratory consultations more than doubled during the winter months.[7,22] Respiratory consultations classified as 'other' had a different pattern with a peak in spring, consistent with seasonal allergies being the primary contributor to this classification group. The high presentation

rates of wheeze-related illness highlights the importance of these conditions in primary care management, and aligns with the high community burden of wheeze identified from other cohort studies.[8,23] The prevalence of otitis media is consistent with other studies.[24]

Childhood respiratory conditions feature highly as a cause of hospital admissions thought to be amenable to preventive activity in primary care. These data suggests only small numbers of children are hospitalised compared to the high volume of respiratory conditions managed within general practice. [10,25] It is possible that paediatric hospitalisations thus represent appropriate care for children with severe respiratory illness, or significant socioeconomic difficulty rather than reflecting unmet need within primary care.

These data also provide information about consulting patterns across across the childhood life course, highlighting the frequency of consultation in the early years and in particular during the first two years of life. While high consultation frequency in the earlier years has been recognised previously, data have usually been grouped within a birth to five years age band,[26] or focused on a single year of life.[27] This study highlights the degree of primary care contact children have in their first two years of life. Strategic management of clinical contact during this time may improve care delivery and enable a balance between preventive and acute care activity.

Strengths and limitations of study

This study examined a very large data set of child-GP consultations including clinical consultation notes, diagnostic codes and prescribing information by way of a software inference algorithm which performed with similar accuracy to clinical experts.[17] The algorithm was designed to maximise specificity, thereby generating a conservative estimate of the burden of childhood respiratory disease in primary care by keeping false positives to a minimum. The presentation and burden of childhood respiratory diseases in primary care has not previously been estimated with such a high degree of accuracy.

Computer algorithms using natural language processing have previously been found to be considerably more accurate than relying on diagnostic codes to make respiratory diagnoses.[28]

Data representing 75 per cent of the child population enrolled within two large primary health organisations, were analysed. The study data set included over 650,000 consultation records (representing over 260,000 person years of data) and the age, ethnic, and socioeconomic characteristics of children enrolled within participating practices were almost identical to those of children enrolled in practices which declined, and to the broader New Zealand population.

This study analysed normal hours primary care GP consultations. The exclusion of nurse-only and out-of-hours consultations may result in an underestimation of primary care respiratory presentation rates. Nurse-only consultations were excluded because only a small proportion of nursing records relate to direct clinical consultations and it was not possible for the algorithm to distinguish these from non-clinical records such as telephone calls.[17] The data set excluded out-of-hours consultations because out-of-hours care is also provided elsewhere to children from consenting practices, consequently PHO out-of-hours data were incomplete.

Although validation of the software algorithm against the gold standard of two expert clinicians' opinion indicated that it had excellent accuracy, particularly with respect to classification of

consultations as respiratory or non-respiratory, this methodology can only provide an estimation of the presentation of these respiratory conditions and resultant service utilisation. It would be impractical to manually check the several hundred thousand consultation records included in the full data set. Notwithstanding this, it is debateable whether manual record review would generate a more accurate estimation.[29,30]

The study used the treating GP's stated diagnosis, or experts' assessment of the presumed diagnosis based upon clinical information and prescriptions recorded, as the gold standard. Consequently, this gold standard included potentially erroneous diagnoses made by the treating GPs,[31,32] and is limited by the information which the GPs determined was pertinent to record. However, the goal of this study was to estimate the burden of illness within primary care as defined by the care received. The GP perception of the conditions being managed is of prime importance in determining health service utilisation and hence this limitation does not affect the algorithm's ability to provide important and useful data.

The need to have conditions with sufficient prevalence to train the algorithm meant that a number of important but less prevalent conditions (e.g. croup, pertussis, and pneumonia) were not able to be individually classified. As a result, the study cannot give estimations of the burden of some diseases, which although relatively rare have considerable morbidity. The algorithm was not designed to differentiate between types of wheeze-related illness given the variation and debate among clinicians regarding the classification of wheeze presentations fpr younger children.[33]

Conclusions and policy implications

These data have demonstrated a clear and consistent pattern in general practice utilisation for children with respiratory illness. Results of this type can assist with general practice workforce planning, and inform debate about current presentation and triage models seen in primary care. The study also highlighted the burden of respiratory disease carried by the youngest members of society and reinforces calls to focus prevention and health promotion campaigns on early stages of the maternal and child health continuum.

The methodology used can be applied to provide similar estimates of respiratory and other conditions and workload across an entire population at all ages. The use of 'big data' in this way also provides a tool for health service planning in primary care which would have increasing application across a wide range of countries.

Footnotes

Acknowledgements

The authors gratefully acknowledge the primary care practices which consented to their consultation records being included in the study dataset, and the primary health organisations which permitted use of proprietary software and resources.

Author contributions

AD, JM, MS, LM, and NT conceived of the study. All authors contributed to the development of the overall study methodology. AD, LM and NT provided clinical input into the algorithm design. JM designed and built the natural language processing tools. JM programmed and trained the algorithm. AD and LM classified the consultation records in the gold standard sets. BD and AD were the principal writers of the manuscript. All authors reviewed and revised the manuscript and approved its final version.

All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Competing interests

All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare support from New Zealand Lotteries Health Research for the submitted work. LM is a director of Compass Health Wellington Trust that might have an interest in the submitted work, no other relationships or activities could appear to have influenced the submitted work

Licence

Funding statement

This work was supported by a New Zealand Lotteries Health Research Grant. The funding body had no role in the collection or analysis of data or the preparation of this manuscript.

Provenance and peer review

Not commissioned; externally peer reviewed.

Data sharing statement

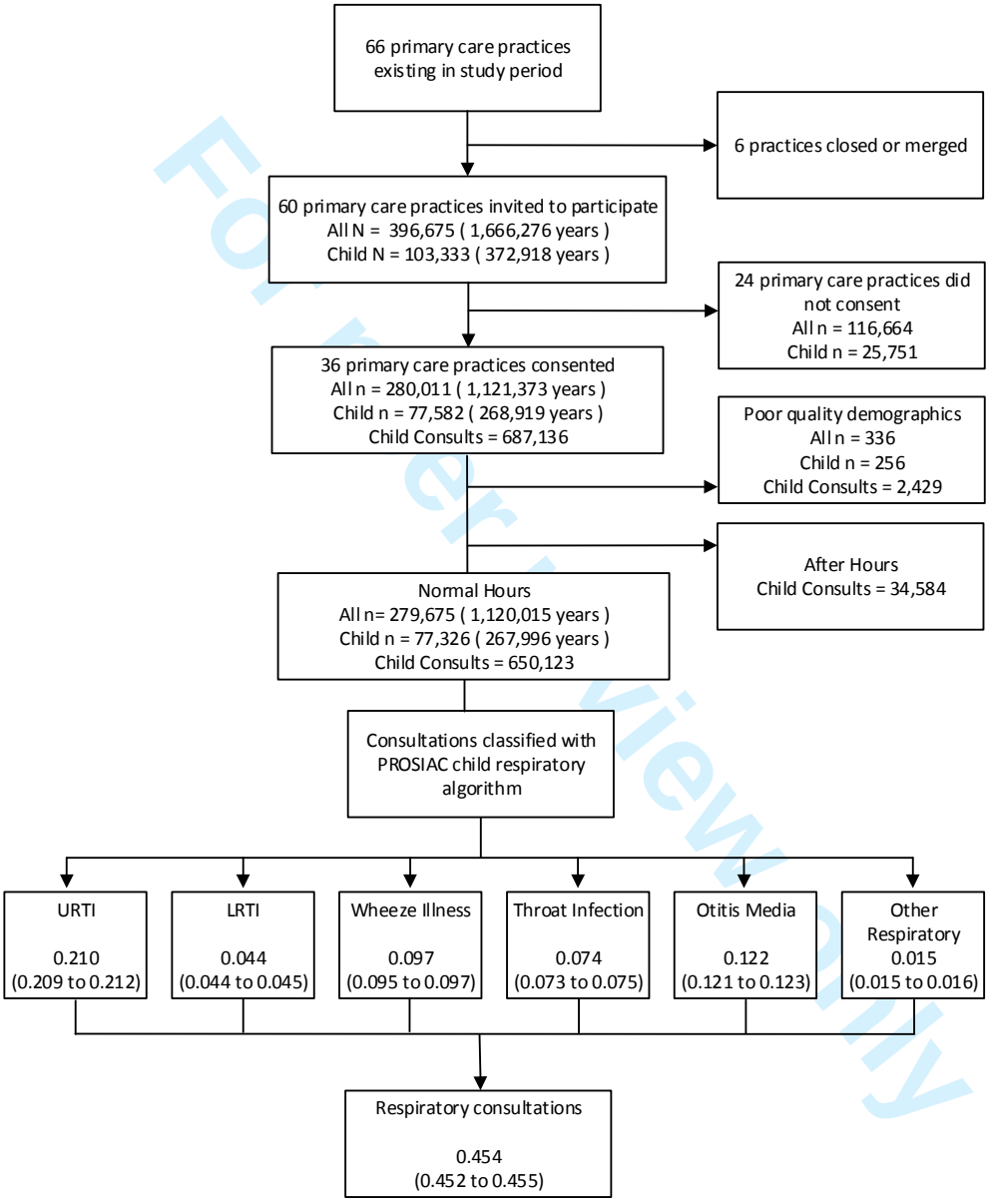
No additional data are available.

Open Access

Ethical approval

This study was approved by the University of Otago Ethics Committee (H13/044)

Figure 1. Selection of child-GP consultation notes and results from analysis.
More than one respiratory condition can be classified in each consultation.
GP = general practitioner; URTI = upper respiratory tract infection; LRTI = lower respiratory tract infection; Wheeze-ill = wheeze-related illness



Appendix 1

Respiratory classification categories and the conditions included in each

Classification category	Respiratory conditions included within category*
Upper respiratory tract infections	• Cold • Croup • Influenza-like illness • Viral influenza in the absence of associated signs or symptoms indicative of lower respiratory tract infection • Scarlet fever • Tracheitis • Cough in the absence of associated signs or symptoms indicative of asthma or lower respiratory tract infection
Lower respiratory tract infections	• Bronchitis • Bronchopneumonia • Chest infection • Chronic lung disease • Cystic fibrosis • Lung abscess/bronchiectasis • Pertussis • Pleurisy • Pneumonia • Tuberculosis • Whooping cough
Wheeze-related illness	• Bronchiolitis • Virus-induced transient wheeze • Persistent wheeze (nonatopic or atopic) • Asthma
Throat infections	• Infectious mononucleosis • Laryngitis • Pharyngitis • Pharyngotonsillitis • Tonsillitis
Otitis media	• Acute otitis media • Chronic suppurative otitis media • Otitis media with effusion • Glue ear

Other respiratory	• Conditions with very low prevalence) for which there are not individual categories
	  ○ Allergic rhinitis
	  ○ Hay fever
	  ○ Rhinitis
	  ○ Sinusitis
	• Consultations in which respiratory symptoms are present but there is insufficient GP entered data to enable classification
	• Consultations in which respiratory symptoms are present with sufficient GP entered data to enable classification but the algorithm fails to classify the consultation

*These classifications are based purely on the information within the electronic health record including consultation notes, medications prescribed and diagnostic Read Codes created on the day of the consultation. It does not include subsequent laboratory tests

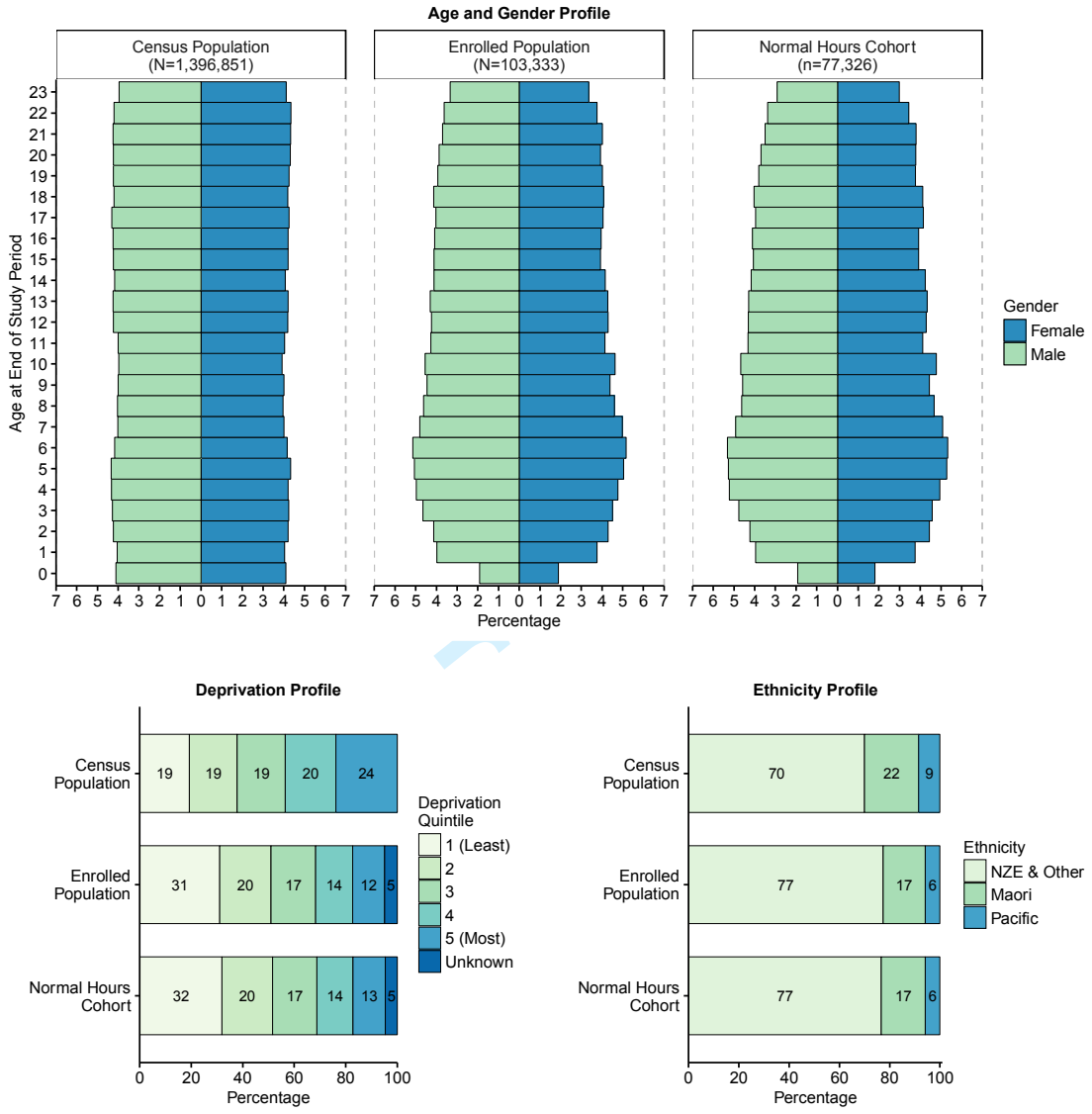
Appendix 2

Automated software inference algorithm measures of performance in the validation set (Set 3)

Diagnostic category	Incidence (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Positive predictive value (95% CI)	Negative predictive value (95% CI)	F-measure (95% CI)
Respiratory	0·46 (0·42 to 0·50)	0·72 (0·67 to 0·78)	0.95 (0.93 to 0.98)	0.93 (0.89 to 0.97)	0.80 (0.76 to 0.84)	0.81 (0.77 to 0.85)
LRTI	0.04 (0.02 to 0.06)	0.61 (0.39 to 0.83)	0.99 (0.98 to 1.00)	0.76 (0.55 to 0.95)	0.98 (0.97 to 0.99)	0.67 (0.47 to 0.85)
URTI	0.21 (0.18 to 0.25)	0.54 (0.45 to 0.64)	0.98 (0.96 to 0.99)	0.86 (0.78 to 0.94)	0.89 (0.86 to 0.92)	0.66 (0.57 to 0.74)
Wheeze ill	0.09 (0.06 to 0.12)	0.96 (0.90 to 1.00)	0.96 (0.94 to 0.98)	0.70 (0.59 to 0.82)	1.00 (0.99 to 1.00)	0.81 (0.73 to 0.89)
Throat infections	0.10 (0.08 to 0.13)	0.50 (0.37 to 0.64)	0.99 (0.99 to 1.00)	0.91 (0.79 to 1.00)	0.95 (0.92 to 0.96)	0.64 (0.51 to 0.76)
Otitis media	0.12 (0.10 to,0.15)	0.58 (0.45 to 0.71)	0.99 (0.98 to 1.00)	0.90 (0.81 to 1.00)	0.94 (0.92 to 0.96)	0.71 (0.59 to 0.81)
Other	0.02 (0.01 to 0.04)	0.66 (0.38 to 0.92)	0.99 (0.98 to 1.00)	0.68 (0.40 to 1.00)	0.99 (0.98 to 1.00)	0.66 (0.42 to 0.87)

LRTI = lower respiratory tract infections; URTI = upper respiratory tract infections; Wheeze ill = wheeze-related illness; other = other respiratory condition.

Appendix 3: Demographic characteristics of children in the study cohort at the end of the study compared with enrolled population and national census data.



Note that the age range extends to 22 because people who were 17 at the start of the study period were 23 at the end of the study. Data stopped being collected from these participants when they turn 18.

The cohort appears to have a low proportion of children under one year of age because the cohort demographic data represents the last day of the study, so many children who entered the cohort under one year of age had moved into a higher age band.

Deprivation quintile is unknown for 0.05 per cent of the census population. This is not shown on the graph.

References

1. Hertzman C, Siddiqi A, Hertzman E, et al. Tackling inequality: get them while they're young. *BMJ* 2010; **340**(7742): 346-8.
2. Lynch JW, Kaplan GA, Cohen RD, et al. Childhood and adult socioeconomic status as predictors of mortality in Finland. *The Lancet* 1994; **343**(8896): 524-7.
3. Marmot MG, Allen JL, Goldblatt P, et al. Fair society, healthy lives: Strategic review of health inequalities in England post-2010. 2010.
4. Walker SP, Wachs TD, Grantham-McGregor S, et al. Inequality in early childhood: risk and protective factors for early child development. *The Lancet* 2011; **378**(9799): 1325-38.
5. The Green Paper for Vulnerable Children. Every child thrives, belongs, achieves. Wellington, NZ.: New Zealand Government, 2011.
6. Ministry of Health. New Zealand Health Survey: Annual update of key findings 2012/13. Wellington, NZ: Ministry of Health, 2013.
7. Chen Y, Kirk M. Incidence of acute respiratory infections in Australia. *Epidemiology and infection* 2014; **142**(07): 1355-61.
8. Asher MI, Montefort S, Björkstén B, et al. Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. *The Lancet* 2006; **368**(9537): 733-43.
9. Gribben B, Salkeld LJ, Hoare S, Jones HF. The incidence of acute otitis media in New Zealand children under five years of age in the primary care setting. *J Prim Health Care* 2012; **4**(3): 205-12.
10. Craig E, Anderson P, Jackson G, Jackson C. Measuring potentially avoidable and ambulatory care sensitive hospitalisations in New zealand children using a newly developed tool. *Journal of the New Zealand Medical Association* 2012; **125**(1366).
11. Davis P, Suaalii-Sauni T, Lay-Yee R, Pearson J. Pacific Patterns in Primary Health Care: A comparison of Pacific and all patient visits to doctors: The National Primary Medical Care Survey (NatMedCa): 2001/02. Wellington: Ministry of Health, 2005.
12. Britt H, Britt H, Miller G, et al. General Practice Activity in Australia 2011-12: BEACH, Bettering the Evaluation And Care of Health: Sydney University Press; 2012.
13. Telfar Barnard L, Baker M, Pierse N, Zhang J. The impact of respiratory disease in New Zealand: 2014 update. *Wellington: The Asthma Foundation* 2015.
14. Dowell A, Turner N. Child health indicators: from theoretical frameworks to practical reality? *Br J Gen Pract* 2014; **64**(629): 608-9.
15. Gill PJ, Goldacre MJ, Mant D, et al. Increase in emergency admissions to hospital for children aged under 15 in England, 1999–2010: national database analysis. *Arch Dis Child* 2013: archdischild-2012-302383.
16. Hobbs FR, Bankhead C, Mukhtar T, et al. Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007–14. *The Lancet* 2016.
17. MacRae J, Darlow B, McBain L, et al. Accessing primary care Big Data: the development of a software algorithm to explore the rich content of consultation records. *BMJ Open* 2015; **5**

.

18. MacRae J, Love T, Baker MG, et al. Identifying influenza-like illness presentation from unstructured general practice clinical narrative using a text classifier rule-based expert system versus a clinical expert. *BMC medical informatics and decision making* 2015; **15**(1): 1.
19. Salmond C, Crampton P, King P, Waldegrave C. NZiDep: a New Zealand index of socioeconomic deprivation for individuals. *Soc Sci Med* 2006; **62**(6): 1474-85.
20. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol* 1978; **107**(1): 71-6.
21. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0; 2014.

22. Chen Y, Williams E, Kirk M. Risk Factors for Acute Respiratory Infection in the Australian Community. 2014.
23. Taussig LM, Wright AL, Holberg CJ, Halonen M, Morgan WJ, Martinez FD. Tucson children's respiratory study: 1980 to present. *Journal of Allergy and Clinical Immunology* 2003; **111**(4): 661-75.
24. Mahadevan M, Navarro-Locsin G, Tan H, et al. A review of the burden of disease due to otitis media in the Asia-Pacific. *Int J Pediatr Otorhinolaryngol* 2012; **76**(5): 623-35.
25. Craig E, Adams JA, Oben G, Reddington A, Wicken A, Simpson JA. The health status of children and young people in the Hutt Valley and Capital and Coast DHBs. Dunedin, New Zealand: NZ Child and Youth Epidemiology Service, 2014.
26. Saxena S, Majeed A, Jones M. Socioeconomic differences in childhood consultation rates in general practice in England and Wales: prospective cohort study. *BMJ* 1999; **318**(7184): 642-6.
27. Golenko XA, Shibl R, Scuffham PA, Cameron CM. Relationship between socioeconomic status and general practitioner visits for children in the first 12 months of life: an Australian study. *Australian Health Review* 2015; **39**(2): 136-45.
28. 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.
29. McColm D, Karcz A. Comparing manual and automated coding of physicians quality reporting initiative measures in an ambulatory EHR. *J Med Pract Manag* 2010; **26**(1): 6-12.
30. Gorelick MH, Knight S, Alessandrini EA, et al. Lack of agreement in pediatric emergency department discharge diagnoses from clinical and administrative data sources. *Acad Emerg Med* 2007; **14**(7): 646-52.
31. Peabody JW, Luck J, Jain S, Bertenthal D, Glassman P. Assessing the accuracy of administrative data in health information systems. *Med Care* 2004; **42**(11): 1066-72.
32. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005; **40**(5p2): 1620-39.
33. Brand PL, Baraldi E, Bisgaard H, et al. Definition, assessment and treatment of wheezing disorders in preschool children: an evidence-based approach. *Eur Respir J* 2008; **32**(4): 1096-110.

What this paper adds

What is already known on this subject

- Previous national and international surveys describe a high estimated annual incidence of respiratory illness but very few data exist to describe the burden of respiratory illness in the community.
- More accurate assessment of illness presentation and service utilisation could be obtained by analysing consultation notes within electronic medical records (EMR) but there are difficulties extracting and analysing the structured and unstructured (free text) data available.
- The lack of primary care data significantly compromises effective planning of primary care services and integration of primary and secondary care activity.

What this study adds

- The development of a natural language processing algorithm enabled the exploration of both structured and unstructured consultation notes.
- 46 per cent of all child-GP consultations are due to presentation of respiratory illness with a remarkably stable year on year pattern of seasonal peaks, quantifying the very high volume of childhood respiratory consultations and workload in general practice, especially during the first two years of life.
- The use of 'big data' methods can help clinicians and planners to refine policy and management in this area, and the method has application to other areas of clinical practice.

BMJ Open

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.

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-017146.R1
Article Type:	Research
Date Submitted by the Author:	23-Jun-2017
Complete List of Authors:	Dowell, Anthony; University of Otago - Wellington , Primary Health Care and General Practice Darlow, Ben; University of Otago, Primary Health Care and General Practice MacRae, Jayden; Datacraft Analytics Stubbe, Maria; University of Otago, Primary Health Care and General Practice Turner, Nikki; University of Auckland, General Practice and Primary Health Care McBain, Lynn; University of Otago, Wellington,
Primary Subject Heading:	General practice / Family practice
Secondary Subject Heading:	Health informatics, Paediatrics, Respiratory medicine
Keywords:	PRIMARY CARE, General Practice, Child health, Big Data, Respiratory illness

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.

Anthony Dowell. MBChB. – corresponding author.
Department of Primary Health Care and General Practice.
University of Otago – Wellington.
23 Mein St, Newtown, Wellington 6242. New Zealand.
Tony.dowell@otago.ac.nz
+64 21 270 1617.

Ben Darlow. PhD
Department of Primary Health Care and General Practice.
University of Otago – Wellington. 23 Mein St, Newtown, Wellington 6242. New Zealand.

Jayden Macrae. MsC
Patients First, Level 4, 50 Customhouse Quay, Wellington 6011, New Zealand.

Maria Stubbe. PhD.
Department of Primary Health Care and General Practice.
University of Otago – Wellington. 23 Mein St, Newtown, Wellington 6242. New Zealand.

Nikki Turner. MD.
Department of General Practice and Primary Health Care. University of Auckland. Level 8, Petherick Towers. 38 Waring Taylor St. Wellington CBD. New Zealand.

Lynn McBain. MD.
Department of Primary Health Care and General Practice.
University of Otago – Wellington. 23 Mein St, Newtown, Wellington 6242. New Zealand.

Word count 3158

Key words
Primary Care, General Practice, Childhood respiratory illness, Natural language software programming, Big data.

Abstract

Objectives.

To identify childhood respiratory tract related illness presentation rates and service utilisation in primary care, by interrogating free text and coded data from Electronic Medical Records.

Design.

Retrospective cohort study. Data interrogation used a natural language processing software inference algorithm.

Setting. 36 primary care practices in New Zealand. Data analysed from January 2008 – December 2013.

Partcipants

The records from 77,582 children enrolled in were reviewed over a six-year period to estimate the presentation of childhood respiratory illness and service utilisation.This cohort represents 268,919 person years of data and over 650,000 unique consultations.

Main outcome measure. Childhood respiratory illness presentation rate to Primary Care practice, with description of seasonal and yearly variation.

Results.

Respiratory conditions constituted 46 per cent of all child-general practitioner consultations with a stable year on year pattern of seasonal peaks. Upper Respiratory Tract Infection was the most common respiratory category accounting for 21.0% of all childhood consultations, followed by otitis media (12.2%), wheeze-related illness (9.7%), throat infection (7.4 %), and Lower Respiratory Tract Infection (4.4 %). Almost 70 per cent of children presented to their general practitioner with at least one respiratory condition in their first year of life; this reduced to approximately 25 per cent for children aged 10 to 17.

Conclusion.

This is the first study to assess the primary care incidence and service utilisation of childhood respiratory illness in a large primary care cohort by interrrogating Electronic Medical Record free text. The study identified the very high primary care workload related to childhood respiratory illness, especially during the first two years of life. These data can enable more effective planning of health service delivery. The findings and methodology have relevance to many countries, and the use of primary care 'big data' in this way can be applied to other health conditions.

Strengths and limitations of this study

- This study uses a novel and validated natural language processing software inference algorithm to identify childhood respiratory illness presentation rates and service utilisation using primary care Big Data.
- The presentation and burden of childhood respiratory diseases in primary care has not previously been estimated with such a high degree of accuracy.
- The algorithm was designed to maximise specificity, thereby generating a conservative estimate of the burden of childhood respiratory disease in primary care by keeping false positives to a minimum
- The methodology has relevance to many OECD countries, and the use of primary care natural language processing in this way can be applied to other health conditions.
- This study analysed normal hours primary care GP consultations. The exclusion of nurse-only and out-of-hours consultations may result in an underestimation of primary care respiratory presentation rates.

Introduction

Childhood is a crucial period for development and well-being. A healthy start to life reduces adulthood morbidity and enhances participation in society.[1-5] Physical illness is an important risk factor for poor health outcomes.[6] Globally, primary care is utilised by all children,[7] but there is currently little published data c of detailed morbidity and utilisation patterns in community settings.

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. Children under five present up to six times a year with acute respiratory infections[8] and high prevalence rates are noted for asthma[9] and otitis media.[10] Such data are, however, mainly reliant on survey responses and parental report. These reports also lack precision regarding individual respiratory conditions, symptom severity, longitudinal patterns and variance related to age and seasonality. These data are needed to effectively plan primary health care service delivery. More detailed hospitalisation data are available[11]; however, these represent an unknown proportion of all cases and are based on diagnostic coding of uncertain accuracy.

International data suggest that respiratory tract related conditions constitute 20 to 25% of all general practitioner (GP) consultations, with higher rates in those under 25 years.[12,13] These data are based on GP self-report, and accuracy may be limited by the competing demands of reporting, meeting patient needs and practice management tasks. Wide variance has been reported in how GPs describe the reason for encounter.[13]

Improved understanding of primary care childhood respiratory illness presentation could enable more systematic approaches to care and resource allocation, and a context for exploring important social and ethnic variations in hospitalisation rates.[11,14] In the Organization for Economic Co-operation and Development (OECD) countries, conditions such as bronchiolitis, asthma, upper respiratory tract infections, and pneumonia make up over 40 % of Ambulatory Sensitive Hospitalisation (ASH); admissions considered preventable through interventions delivered outside of hospitals, predominantly within primary care.[6,11,15,16]

More accurate assessment of illness presentation and service utilisation could be obtained by analysing consultation notes within electronic medical records (EMR) common in OECD primary care settings. While there has been some exploration of the potential for 'big data' assessment of general practice workload,[17] these data have not previously been used to analyse childhood respiratory service utilisation due to difficulties with extracting and analysing both structured and unstructured data available (primarily clinical consultation notes). The development of novel software has enabled the exploration of New Zealand EMR data.[18,19]

This study aimed to interrogate data from EMR to identify primary care presentation and service utilisation related to common childhood respiratory tract conditions and their complications.

Methods

Design

A natural language processing software inference algorithm was developed to interrogate quantitative and qualitative cross-sectional and retrospective cohort data from EMR.[18,19]

Setting and participants

Figure 1 illustrates the creation and analysis of the dataset. In New Zealand there is universal enrolment with a primary care practice. As more fully described in an earlier publication outlining the development of the design[17], the study was conducted in the Wellington region of New Zealand, a mixed urban and rural setting. It consisted of 36 consenting practices of 60 in total from two primary health organisations (PHOs). This comprised 75% of the total childhood population under 18 years of age of these combined PHOs. There was a total of 77 467 children enrolled in these practices over the study period between 1 January 2008 until 31 December 2013, including children that both joined and left this cohort during this period. Changes included births, deaths, turning 18 years, or moving into or out of a practice. This cohort represented 268,919 person years.

Data were collected directly from EMR using software which automates the extraction, and secure transmission of large data sets. The dataset comprised records from consultations generated during both standard office hours and out-of-hours practice. Data were extracted from the EMR for all child-GP consultations at consenting practices during the study period (n= 687,136). Each consultation record was connected to an individual's National Health Index (NHI) number. The is a unique identifier assigned to every person who uses health services in New Zealand and enabled records to be matched between datasets. Consultations for which there were poor quality data (2439 consults from 256 children) were excluded. Out-of-hours consultations (n=34,584) were not analysed due to differing participation in out-of-hours services by the practices. All data were analysed within the PHO which has rigorous protocols in place to ensure patient confidentiality. The research team had not access to identifiable data.

Process

Each of the 650,123 clinical consultation notes was interrogated by a software inference algorithm and hierarchical classification system described previously.[19] The algorithm classified consultation records using: clinical information recorded by GPs and practice nurses, any recorded Read code diagnostic classifications, and prescribing information. The first level of the hierarchy divided all consultations into either 'respiratory' or 'not respiratory'. Note that 'respiratory' here included all respiratory tract-related conditions and presentations and the associated complication of otitis media. The 'not respiratory' category included consultations where the primary presentation and diagnosis was for conditions such as injury or gastroenteritis, and consultations in which the respiratory system was examined and screened, but no signs, symptoms, or diagnoses were recorded. These screening consultations were excluded so that the burden of respiratory tract illness estimate was not inflated by consultations which did not result from a respiratory tract illness or its complications

The second level of the hierarchy sub-classified consultations into one or more specific respiratory categories. These categories were determined by a group of clinical experts; consideration was given to the degree to which conditions could be mapped to high prevalence (that which is common)

and/or responsible for significant morbidity and hospitalisation (that which is important).The six categories were i) upper respiratory tract infections (URTI); ii) lower respiratory tract infections (LRTI); iii) wheeze-related illnesses; iv) throat infections; v) otitis media; and vi) other respiratory conditions. The conditions included within each diagnostic category are presented in Appendix 1.

The algorithm was trained, tested, and validated using three independent gold standard data sets of 1200 consultation records which had been independently classified by two general practice clinical experts (AD and LM). The algorithm was designed to replicate the judgements made by these clinical experts. Development aimed to optimise specificity while maximising sensitivity to minimise the occurrence of false positives. The algorithm's sensitivity, specificity, positive and negative predictive values, and F-measure for each of the diagnostic categories against a gold standard validation set of 1200 consultation records has been published previously.[18]

Analysis

The demographic characteristics of age, gender, ethnicity (NZ indigenous Māori, Pacific, other), and New Zealand Deprivation Index (a measure of socioeconomic deprivation[20]) of the cohort (n=77,326) were compared with those of all children enrolled within the two PHOs (N=103,333) and the New Zealand population using national census data.

The proportion of primary care consultations for children aged 18 years and under which were related to the six specific respiratory conditions outlined above was obtained from the dataset using the algorithm. The utilisation of services for these six conditions was analysed by demographic characteristics. Consultation rates are expressed per 1000 child years observed due to the differing length of time individuals might be participants in the cohort. Patients were observed for the period in which they were enrolled in a participating practice; this was calculated from the date of a child's first visit to a practice until they were removed from the enrolment register. Both deprivation and ethnicity status were taken as the last ethnicity and deprivation recorded from the GP records. Consultation rates were adjusted for sensitivity and specificity of the algorithm [18]and a direct standardization method was applied to level 2 ethnicity and socio-economic deprivation quintiles against NZ Census 2013 data. Estimates of true rates were made using final test sensitivity and specificity results for each classification category using the method described by Rogan and Gladen. [21]All Data aggregation, transformation, cleaning and storage was done in Microsoft SQL Server, and statistical analysis was undertaken in R using packages including boot, epiR, combinat, stats, tm, RWeka, slam, SnowballC and caret. [22]
STROBE Guidelines were followed.

Results

The demographic characteristics of the study cohort closely matched those of the enrolled population (Appendix 2). The age distribution of the study cohort also closely matched the national comparison data. Compared with national census data the study cohort had a greater proportion of children from the least deprived quintile grouping (32% vs 25%) and a lower proportion of Māori (17% vs 22%).

From the 650,123 consultations reviewed the true rate of presentation for a respiratory tract condition or complication was calculated to be 45.4 per cent of all consultations for children under 18 years of age (Figure 1). URTI was the most common respiratory tract related category represented in 21.0% (95% CI 20.9-21.2%) of all consultations, followed by otitis media (12.2% CI

12.1-12.3%), wheeze-related illness (9.7% CI 9.5-9.7%), throat infection (7.4 % CI 7.3-7.5%), and LRTI (4.4 % CI 4.4 – 4.5%). Other respiratory tract related classifications accounted for just 1.5% of all consultations . One respiratory tract related condition was classified in 27.6% of all consultations, two in 7.0%, three in 0.8%, and greater than three in 0.1%. The rates of child respiratory tract condition or complication consultation were 1,101 per 1,000 person years observed for all respiratory tract conditions and complications, 509 per 1000 for URTI, 107 per 1000 for LRTI, 235 per 1000 for Wheeze Illness, 180 per 1000 for Throat Infections, 296 per 1000 for Otitis Media and 36 per 1000 for Other Respiratory conditions. The incidence of both respiratory and non-respiratory tract related consultations remained stable throughout the study period with a consistent pattern of seasonal peaks and troughs (Fig 2). The respiratory tract related consultation rate was highest in the Southern Hemisphere winter month of August, and lowest in January (Figure 3). Non-respiratory consultations followed a similar pattern but with shallower peaks and troughs. Respiratory tract related conditions explained 64.4% of the annual seasonal variation in child consultation rates. All respiratory tract related conditions which were sub-classified followed a similar pattern of peaks in August and troughs in January except for 'other' respiratory tract conditions which were highest in December and lowest in April (Figure 4). Figures 4 and 5 present the annual variation in respiratory tract condition related presentation for each classification category.

Respiratory tract related consultations occurred throughout childhood, but at much greater frequency during the first two years of life. (Figure 6) During the first year of life 73.5 % of children presented to their GP with at least one respiratory tract related condition. Following the second year of life, the presentation of all respiratory tract related conditions decreased with increasing age. Of children aged 10 to 17 years, 22.5 % presented with at least one respiratory condition (Figure 6). 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.

Fig 2 : Respiratory tract related and non-respiratory consultations per quarter per 1000 enrolled children January 2008 to December 2013.

Figure 3: Mean respiratory tract related and non-respiratory consultations per month per 1000 enrolled children. January 2008 to December 2013 demonstrating seasonal variation.

Figure 4: Yearly variation of consultations per month per 1000 enrolled children for each respiratory tract related illness category - January 2008 to December 2013.

Figure 5 : Mean consultations per month per 1000 enrolled children for each respiratory tract related illness category - January 2008 to December 2013

Figure 6: Respiratory tract related consultation frequency by selected age cohort.

Figure 6: Each facet includes children who were enrolled within that age band for a twelve month period (e.g. from the day they turned one until the day before they turned two). The cohort of children under one is small because many children do not enrol until they are over three months old and may therefore only be enrolled for nine months before turning one.

Discussion

This is the first study to assess the primary care incidence and service utilisation of childhood respiratory tract related illness in such a large cohort observing over 250,000 person years and more than 650,000 unique consultations. Using a novel and validated method of interrogating EMR free text, this study found that respiratory tract related conditions constituted 45.4 per cent of all child-GP consultations. This quantifies the very high volume of childhood respiratory tract related consultations and workload in general practice, especially during the first two years of life. These data can enable more effective planning of primary care service delivery and indicate areas in which to focus preventive programmes. The study also highlights the high presentation rates to primary care of those respiratory conditions which frequently present for hospital admission.

Comparison with other studies

The presentation rate of respiratory illness and pattern of seasonal peaks was remarkably stable across the six years included in this data set and was unchanged by events such as the H1N1 influenza pandemic of 2009. Consistent with findings from Australian[8,23,] and Chilean surveys [24], the presentation of nearly all respiratory consultations more than doubled during the winter months, and providing a comparator with 'seasonal ' changes between wet and dry seasons seen in tropical regions. [25] Respiratory consultations classified as 'other' had a different pattern with a peak in spring, consistent with seasonal allergies being the primary contributor to this classification group. The high presentation rates of wheeze-related illness highlights the importance of these conditions in primary care management, and aligns with the high community burden of wheeze identified from other cohort studies.[9,26] The prevalence of otitis media is consistent with other studies.[27]

While we could find no other studies that used an NLP methodology, our findings are consistent with the stated underestimate of respiratory illness prevalence reported from a recent primary care study in Ireland using Read code data only. [28]

Childhood respiratory conditions feature highly as a cause of hospital admissions thought to be amenable to preventive activity in primary care. These data suggests only small numbers of children are hospitalised compared to the high volume of respiratory conditions managed within general practice. [11,29] It is possible that paediatric hospitalisations thus represent appropriate care for children with severe respiratory illness, or significant socioeconomic difficulty rather than reflecting unmet need within primary care.

These data also provide information about consulting patterns across across the childhood life course, highlighting the frequency of consultation in the early years and in particular during the first two years of life. While high consultation frequency in the earlier years has been recognised previously, data have usually been grouped within a birth to five years age band,[30] or focused on a

single year of life.[31] This study highlights the degree of primary care contact children have in their first two years of life. Strategic management of clinical contact during this time may improve care delivery and enable a balance between preventive and acute care activity.

Strengths and limitations of study

This study examined a very large data set of child-GP consultations including clinical consultation notes, diagnostic codes and prescribing information by way of a software inference algorithm which performed with similar accuracy to clinical experts.[18] The algorithm was designed to maximise specificity, thereby generating a conservative estimate of the burden of childhood respiratory disease in primary care by keeping false positives to a minimum. The presentation and burden of childhood respiratory diseases in primary care has not previously been estimated with such a high degree of accuracy.

Computer algorithms using natural language processing have previously been found to be considerably more accurate than relying on diagnostic codes to make respiratory diagnoses.[32]

Data representing 75 per cent of the child population enrolled within two large primary health organisations, were analysed. The study data set included over 650,000 consultation records (representing over 260,000 person years of data) and the age, ethnic, and socioeconomic characteristics of children enrolled within participating practices were almost identical to those of children enrolled in practices which declined, and to the broader New Zealand population.

This study analysed normal hours primary care GP consultations. The exclusion of nurse-only and out-of-hours consultations may result in an underestimation of primary care respiratory tract related presentation rates. Nurse-only consultations were excluded because only a small proportion of nursing records relate to direct clinical consultations and it was not possible for the algorithm to distinguish these from non-clinical records such as telephone calls.[18] The data set excluded out-of-hours consultations because out-of-hours care is also provided elsewhere to children from consenting practices, consequently PHO out-of-hours data were incomplete.

Although validation of the software algorithm against the gold standard of two expert clinicians' opinion indicated that it had excellent accuracy, particularly with respect to classification of consultations as respiratory tract related or non-respiratory, this methodology can only provide an estimation of the presentation of these respiratory conditions and resultant service utilisation. It would be impractical to manually check the several hundred thousand consultation records included in the full data set. Notwithstanding this, it is debateable whether manual record review would generate a more accurate estimation.[33,34]

The gold standard used for this study was the GP's stated diagnosis, matched to GP experts' assessment based on clinical data available. There is the potential for error in the GP decision-making,[35,36] and is limited by the amount and detail of the recorded information by each GP. While recognising this limitation accuracy of GP diagnosis was not the prime purpose of this study, the intention was to estimate illness and health service utilisation as identified by the GP records.

The algorithm requires common conditions with sufficient prevalence to allow effective training. Therefore some important but less prevalent conditions (e.g. croup, pertussis, and pneumonia) required to be grouped. As a result, the study cannot give estimations of the burden of some

diseases, which although relatively rare have considerable morbidity. The algorithm was not designed to differentiate between types of wheeze-related illness given the variation and debate among clinicians regarding the classification of wheeze presentations fpr younger children.[37]

Conclusions and policy implications

These data have demonstrated a clear and consistent pattern in general practice utilisation for children with respiratory tract related illness. Results of this type can assist with general practice workforce planning, and inform debate about current presentation and triage models seen in primary care. The study also highlighted the burden of respiratory disease carried by the youngest members of society and reinforces calls to focus prevention and health promotion campaigns on early stages of the maternal and child health continuum.

The methodology used can be applied to provide similar estimates of respiratory and other conditions and workload across an entire population at all ages. The use of NLP software in this way also provides a tool for health service planning in primary care which would have increasing application across a wide range of countries.

Footnotes

Acknowledgements

The authors gratefully acknowledge the primary care practices which consented to their consultation records being included in the study dataset, and the primary health organisations which permitted use of proprietary software and resources.

Author contributions

AD, JM, MS, LM, and NT conceived of the study. All authors contributed to the development of the overall study methodology. AD, LM and NT provided clinical input into the algorithm design. JM designed and built the natural language processing tools. JM programmed and trained the algorithm. AD and LM classified the consultation records in the gold standard sets. BD and AD were the principal writers of the manuscript. All authors reviewed and revised the manuscript and approved its final version.

All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Competing interests

All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare support from New Zealand Lotteries Health Research for the submitted work. LM is a director of

Compass Health Wellington Trust that might have an interest in the submitted work, no other relationships or activities could appear to have influenced the submitted work

Licence

Funding statement

This work was supported by a New Zealand Lotteries Health Research Grant. The funding body had no role in the collection or analysis of data or the preparation of this manuscript.

Provenance and peer review

Not commissioned; externally peer reviewed.

Data sharing statement

No additional data are available.

Open Access

Ethical approval

This study was approved by the University of Otago Ethics Committee (H13/044)

Figure 1. Selection of child-GP consultation notes and results from analysis.
More than one respiratory condition can be classified in each consultation.
GP = general practitioner; URTI = upper respiratory tract infection; LRTI = lower respiratory tract infection; Wheeze-ill = wheeze-related illness

Appendix 1

Appendix 2

References

1. Hertzman C, Siddiqi A, Hertzman E, et al. Tackling inequality: get them while they're young. *BMJ* 2010; **340**(7742): 346-8.
2. Lynch JW, Kaplan GA, Cohen RD, et al. Childhood and adult socioeconomic status as predictors of mortality in Finland. *The Lancet* 1994; **343**(8896): 524-7.
3. Marmot MG, Allen JL, Goldblatt P, et al. Fair society, healthy lives: Strategic review of health inequalities in England post-2010. 2010.
4. Walker SP, Wachs TD, Grantham-McGregor S, et al. Inequality in early childhood: risk and protective factors for early child development. *The Lancet* 2011; **378**(9799): 1325-38.
5. Bethell CD, Newacheck P, Hawes E, Halfon N. Adverse childhood experiences: assessing the impact on health and school engagement and the mitigating role of resilience. *Health Affairs* 2014; **33**(12): 2106-15.
6. The Green Paper for Vulnerable Children. Every child thrives, belongs, achieves. Wellington, NZ.: New Zealand Government, 2011.
7. Ministry of Health. New Zealand Health Survey: Annual update of key findings 2012/13. Wellington, NZ: Ministry of Health, 2013.
8. Chen Y, Kirk M. Incidence of acute respiratory infections in Australia. *Epidemiology and infection* 2014; **142**(07): 1355-61.
9. Asher MI, Montefort S, Björkstén B, et al. Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. *The Lancet* 2006; **368**(9537): 733-43.
10. Gribben B, Salkeld LJ, Hoare S, Jones HF. The incidence of acute otitis media in New Zealand children under five years of age in the primary care setting. *J Prim Health Care* 2012; **4**(3): 205-12.
11. Craig E, Anderson P, Jackson G, Jackson C. Measuring potentially avoidable and ambulatory care sensitive hospitalisations in New zealand children using a newly developed tool. *Journal of the New Zealand Medical Association* 2012; **125**(1366).
12. Davis P, Suaalii-Sauni T, Lay-Yee R, Pearson J. Pacific Patterns in Primary Health Care: A comparison of Pacific and all patient visits to doctors: The National Primary Medical Care Survey (NatMedCa): 2001/02. Wellington: Ministry of Health, 2005.
13. Britt H, Britt H, Miller G, et al. General Practice Activity in Australia 2011-12: BEACH, Bettering the Evaluation And Care of Health: Sydney University Press; 2012.
14. Telfar Barnard L, Baker M, Pierse N, Zhang J. The impact of respiratory disease in New Zealand: 2014 update. *Wellington: The Asthma Foundation* 2015.
15. Dowell A, Turner N. Child health indicators: from theoretical frameworks to practical reality? *Br J Gen Pract* 2014; **64**(629): 608-9.
16. Gill PJ, Goldacre MJ, Mant D, et al. Increase in emergency admissions to hospital for children aged under 15 in England, 1999–2010: national database analysis. *Arch Dis Child* 2013: archdischild-2012-302383.
17. Hobbs FR, Bankhead C, Mukhtar T, et al. Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007–14. *The Lancet* 2016.
18. MacRae J, Darlow B, McBain L, et al. Accessing primary care Big Data: the development of a software algorithm to explore the rich content of consultation records. *BMJ Open* 2015; **5**

.

19. MacRae J, Love T, Baker MG, et al. Identifying influenza-like illness presentation from unstructured general practice clinical narrative using a text classifier rule-based expert system versus a clinical expert. *BMC medical informatics and decision making* 2015; **15**(1): 1.
20. Salmond C, Crampton P, King P, Waldegrave C. NZiDep: a New Zealand index of socioeconomic deprivation for individuals. *Soc Sci Med* 2006; **62**(6): 1474-85.
21. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol* 1978; **107**(1): 71-6.

22. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0; 2014.
23. Chen Y, Williams E, Kirk M. Risk Factors for Acute Respiratory Infection in the Australian Community. 2014.
24. Astudillo P, Mancilla P, Olmos C, Reyes Á. Epidemiology of pediatric respiratory consultations in Santiago de Chile, from 1993 to 2009. *Revista Panamericana de Salud Pública* 2012; **32**(1): 56-61.
25. Rosa AM, Ignotti E, Botelho C, Castro HAd, Hacon SdS. Respiratory disease and climatic seasonality in children under 15 years old in a town in the Brazilian Amazon. *J Pediatr (Rio J)* 2008; **84**(6): 543-9.
26. Taussig LM, Wright AL, Holberg CJ, Halonen M, Morgan WJ, Martinez FD. Tucson children's respiratory study: 1980 to present. *Journal of Allergy and Clinical Immunology* 2003; **111**(4): 661-75.
27. Mahadevan M, Navarro-Locsin G, Tan H, et al. A review of the burden of disease due to otitis media in the Asia-Pacific. *Int J Pediatr Otorhinolaryngol* 2012; **76**(5): 623-35.
28. Molony D, Beame C, Behan W, et al. 70,489 primary care encounters: retrospective analysis of morbidity at a primary care centre in Ireland. *Irish Journal of Medical Science (1971-)* 2016; **185**(4): 805-11.
29. Craig E, Adams JA, Oben G, Reddington A, Wicken A, Simpson JA. The health status of children and young people in the Hutt Valley and Capital and Coast DHBs. Dunedin, New Zealand: NZ Child and Youth Epidemiology Service, 2014.
30. Saxena S, Majeed A, Jones M. Socioeconomic differences in childhood consultation rates in general practice in England and Wales: prospective cohort study. *BMJ* 1999; **318**(7184): 642-6.
31. Golenko XA, Shibl R, Scuffham PA, Cameron CM. Relationship between socioeconomic status and general practitioner visits for children in the first 12 months of life: an Australian study. *Australian Health Review* 2015; **39**(2): 136-45.
32. 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.
33. McColm D, Karcz A. Comparing manual and automated coding of physicians quality reporting initiative measures in an ambulatory EHR. *J Med Pract Manag* 2010; **26**(1): 6-12.
34. Gorelick MH, Knight S, Alessandrini EA, et al. Lack of agreement in pediatric emergency department discharge diagnoses from clinical and administrative data sources. *Acad Emerg Med* 2007; **14**(7): 646-52.
35. Peabody JW, Luck J, Jain S, Bertenthal D, Glassman P. Assessing the accuracy of administrative data in health information systems. *Med Care* 2004; **42**(11): 1066-72.
36. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005; **40**(5p2): 1620-39.
37. Brand PL, Baraldi E, Bisgaard H, et al. Definition, assessment and treatment of wheezing disorders in preschool children: an evidence-based approach. *Eur Respir J* 2008; **32**(4): 1096-110.

What this paper adds

What is already known on this subject

- Previous national and international surveys describe a high estimated annual incidence of respiratory illness but very few data exist to describe the burden of respiratory illness in the community.
- More accurate assessment of illness presentation and service utilisation could be obtained by analysing consultation notes within electronic medical records (EMR) but there are difficulties extracting and analysing the structured and unstructured (free text) data available.
- The lack of primary care data significantly compromises effective planning of primary care services and integration of primary and secondary care activity.

What this study adds

- The development of a natural language processing algorithm enabled the exploration of both structured and unstructured consultation notes.
- 46 per cent of all child-GP consultations are due to presentation of respiratory illness with a remarkably stable year on year pattern of seasonal peaks, quantifying the very high volume of childhood respiratory consultations and workload in general practice, especially during the first two years of life.
- The use of 'big data' methods can help clinicians and planners to refine policy and management in this area, and the method has application to other areas of clinical practice.

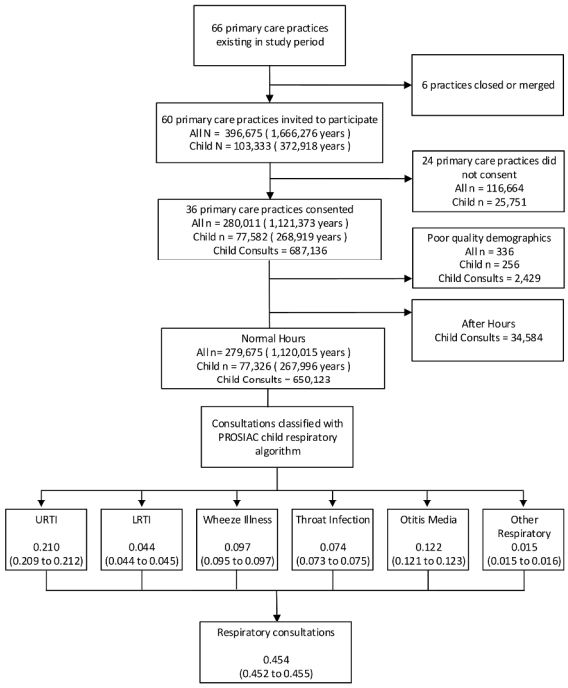


Figure 1. Selection of child-GP consultation notes and results from analysis.

297x420mm (300 x 300 DPI)

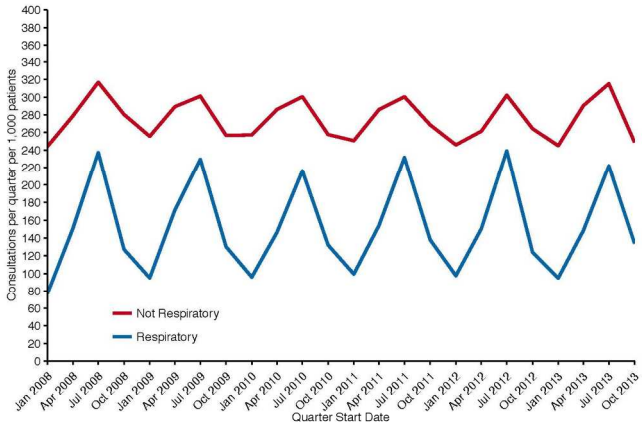


Fig 2 : Respiratory tract related and non-respiratory consultations per quarter per 1000 enrolled children January 2008 to December 2013.

297x420mm (300 x 300 DPI)

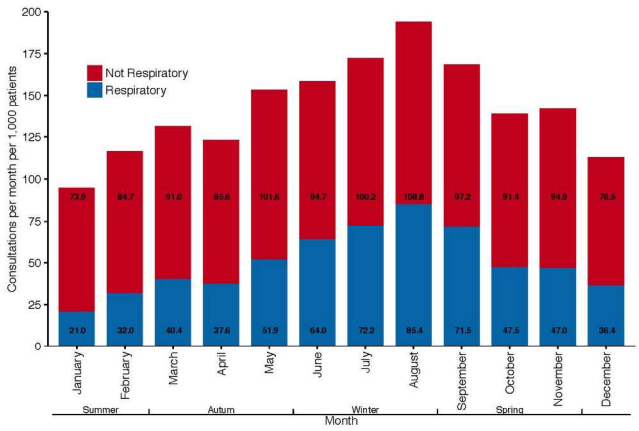


Figure 3: Mean respiratory tract related and non-respiratory consultations per month per 1000 enrolled children. January 2008 to December 2013 demonstrating seasonal variation.

297x420mm (300 x 300 DPI)

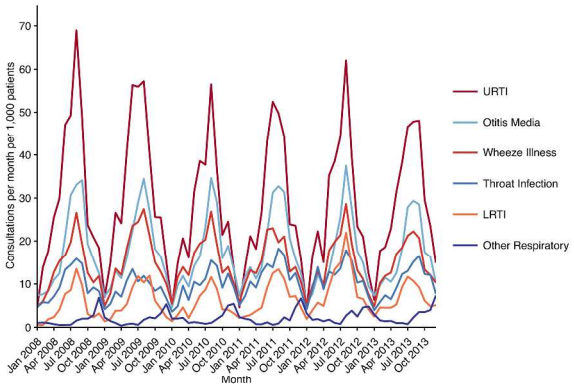


Figure 4: Yearly variation of consultations per month per 1000 enrolled children for each respiratory tract related illness category - January 2008 to December 2013.

297x420mm (300 x 300 DPI)

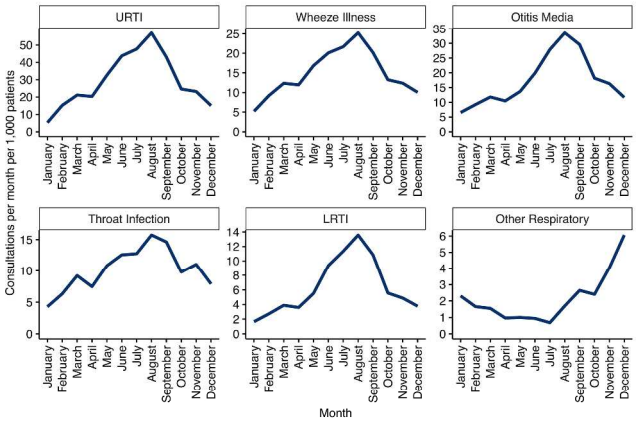


Figure 5 : Mean consultations per month per 1000 enrolled children for each respiratory tract related illness category - January 2008 to December 2013

297x420mm (300 x 300 DPI)

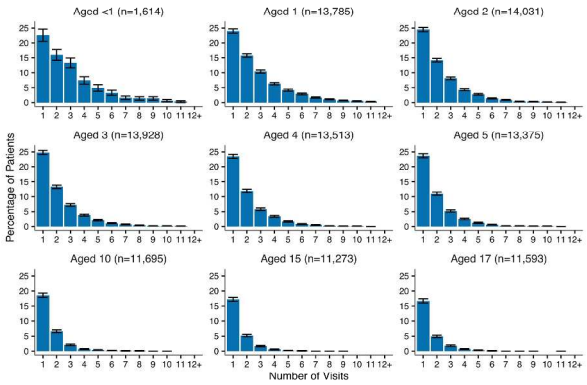


Figure 6: Respiratory tract related consultation frequency by selected age cohort.

297x420mm (300 x 300 DPI)

Appendix 1

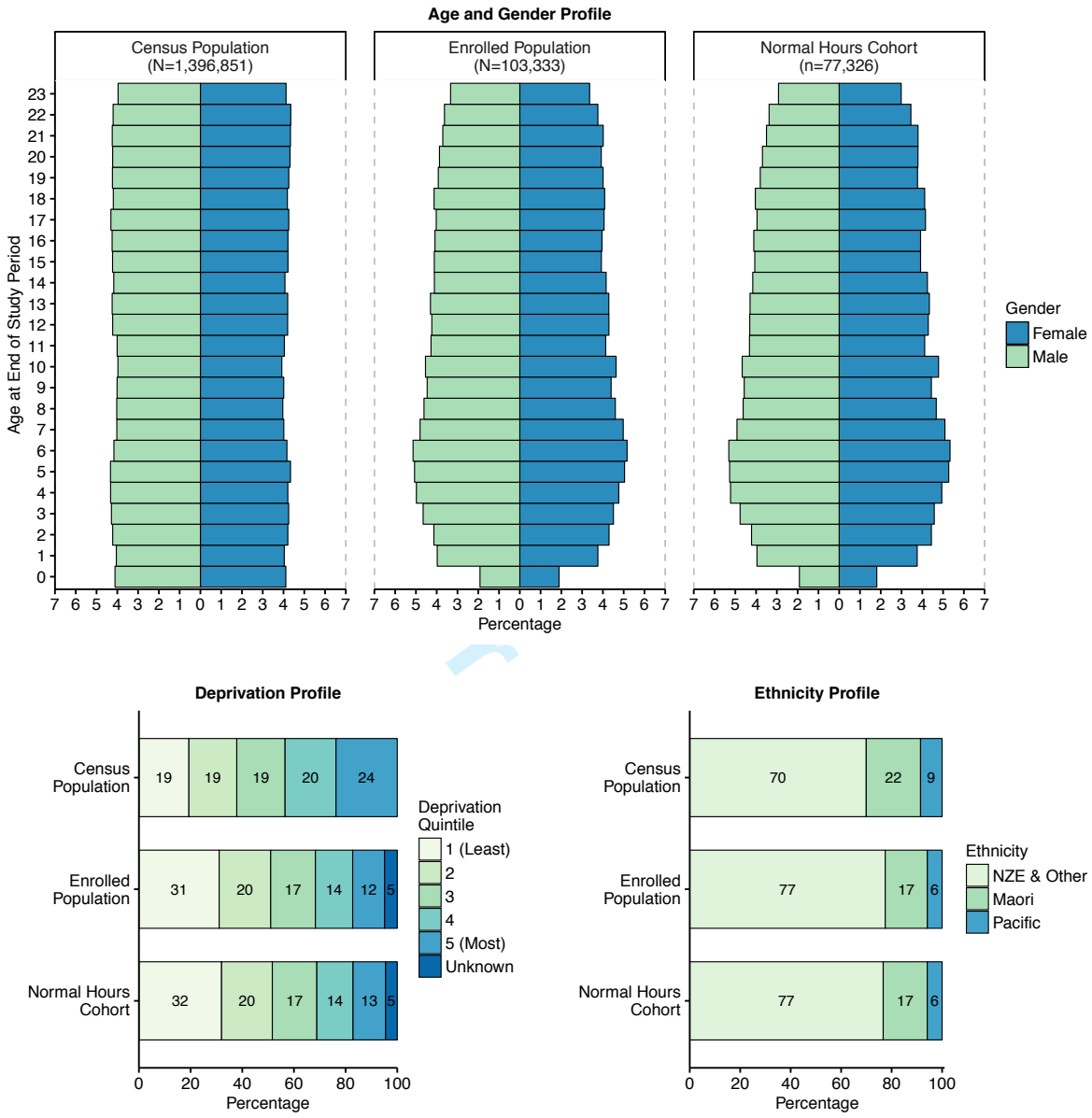
Respiratory classification categories and the conditions included in each

Classification category	Respiratory conditions included within category*
Upper respiratory tract infections	• Cold • Croup • Influenza-like illness • Viral influenza in the absence of associated signs or symptoms indicative of lower respiratory tract infection • Scarlet fever • Tracheitis • Cough in the absence of associated signs or symptoms indicative of asthma or lower respiratory tract infection
Lower respiratory tract infections	• Bronchitis • Bronchopneumonia • Chest infection • Chronic lung disease • Cystic fibrosis • Lung abscess/bronchiectasis • Pertussis • Pleurisy • Pneumonia • Tuberculosis • Whooping cough
Wheeze-related illness	• Bronchiolitis • Virus-induced transient wheeze • Persistent wheeze (nonatopic or atopic) • Asthma
Throat infections	• Infectious mononucleosis • Laryngitis • Pharyngitis • Pharyngotonsillitis • Tonsillitis
Otitis media	• Acute otitis media • Chronic suppurative otitis media • Otitis media with effusion • Glue ear

Other respiratory	• Conditions with very low prevalence) for which there are not individual categories
	  ○ Allergic rhinitis
	  ○ Hay fever
	  ○ Rhinitis
	  ○ Sinusitis
	• Consultations in which respiratory symptoms are present but there is insufficient GP entered data to enable classification
	• Consultations in which respiratory symptoms are present with sufficient GP entered data to enable classification but the algorithm fails to classify the consultation

*These classifications are based purely on the information within the electronic health record including consultation notes, medications prescribed and diagnostic Read Codes created on the day of the consultation. It does not include subsequent laboratory tests

Appendix 2: Demographic characteristics of children in the study cohort at the end of the study compared with enrolled population and national census data.



Note that the age range extends to 22 because people who were 17 at the start of the study period were 23 at the end of the study. Data stopped being collected from these participants when they turn 18.

The cohort appears to have a low proportion of children under one year of age because the cohort demographic data represents the last day of the study, so many children who entered the cohort under one year of age had moved into a higher age band.

Deprivation quintile is unknown for 0.05 per cent of the census population. This is not shown on the graph.

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.

Strobe Checklist of items that should be included in reports of *cohort studies*

	Item	Recommendation	Completed
Title and abstract	1	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found	Yes – Page 1 Yes – Page 1 Lines 34 - 52
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	Yes, - Page 4 Lines 10-46
Objectives	3	State specific objectives, including any prespecified hypotheses	Yes, Page 4 Lines 50-51
Methods			
Study design	4	Present key elements of study design early in the paper	Yes Page 5 Lines 6 – 7
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Yes – Page 5 Lines 10-22
Participants	6	(a) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up (b) For matched studies, give matching criteria and number of exposed and unexposed	Yes Page 5 Lines 16-22 N/A
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect	Yes – Page 5 Line 40 to P

		modifiers. Give diagnostic criteria, if applicable	6 Line 16
Datasources/measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Yes PAGE 6 Line 25-42 N/A
Bias	9	Describe any efforts to address potential sources of bias	Yes Page 5 Line 40-54
Study size	10	Explain how the study size was arrived a	Yes Page 5 Line 16-22
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	Yes Page 6 Line 26 - 39
Statistical methods	12	9a)Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) If applicable, explain how loss to follow-up was addressed (e) Describe any sensitivity analysis	Yes Page 6 Line 26 -39 Yes Page 5 Line 40-60 Yes, Fig 1 Yes, refer to original methodology paper
Results			
Participants	13	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed (b) Give reasons for non-participation at each stage (c) Consider use of a flow diagram	Yes – Fig 1 P15 Yes – Fig 1 Yes – Fig 1
Descriptive data	14	(a)Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential	Yes – Fig 1

		confounders (b) Indicate number of participants with missing data for each variable of interest (c) Summarise follow-up time (eg, average and total amount	 Yes Yes Page 1 Line 10-12
Outcome data	15	Report numbers of outcome events or summary measures over time	Yes Page 6 **Line 53 – P7**
Main results	16	(a)Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included (b) Report category boundaries when continuous variables were categorized (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	Yes Page 6/7 Line 57 - 4 **N/A** N/A
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	YES Page 7 Line 14-33
Discussion			
Key results	18	Summarise key results with reference to study objectives	YES Page 10
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	YES Page 11 Line 52
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	YES Page 11/12
Generalisability	21	Discuss the generalisability (external validity) of the study results	YES Page 12 Line 33-45
Other information			

Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	YES Page 13 Line 52/53