HHS Public Access

Author manuscript

J Allergy Clin Immunol Pract. Author manuscript; available in PMC 2019 January 01.

Published in final edited form as:

J Allergy Clin Immunol Pract. 2018 ; 6(1): 126–131. doi:10.1016/j.jaip.2017.04.041.

Natural Language Processing for Asthma Ascertainment in Different Practice Settings

Chung-Il Wi, MD[1,2,*], **Sunghwan Sohn, PhD**[3,*], **Mir Ali, MD**[4], **Elizabeth Krusemark**[1,2], **Euijung Ryu, PhD**[3], **Hongfang Liu, PhD**[3,¶], and **Young J. Juhn, MD, MPH**[1,2,¶]

[1]Department of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, Minnesota

[2]Asthma Epidemiology Research Unit, Mayo Clinic, Rochester, Minnesota

[3]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota

[4]Department of Pediatrics, Sanford Children's Hospital, Sioux Falls, South Dakota

Abstract

Background—We developed and validated NLP-PAC, a natural language processing (NLP) algorithm based on Predetermined Asthma Criteria (PAC) for asthma ascertainment using electronic health records (EHRs) at Mayo Clinic.

Objective—To adapt NLP-PAC in a different health care setting, Sanford Children Hospital (SCH) by assessing the external validity of NLP-PAC.

Methods—The study was designed as a retrospective cohort study, which utilized a random sample of 2011–2012 Sanford Birth cohort (n=595). Manual chart review was performed on the cohort for asthma ascertainment based on PAC. We then used half of the cohort as a training cohort (n=298) and the other half as a blind test cohort to evaluate the adapted NLP-PAC algorithm. Association of known asthma-related risk factors with the Sanford-NLP algorithm-driven asthma ascertainment was tested.

Results—Among the eligible test cohort (n=297), 160 (53%) were males, 268 (90%) White, and the median age was 2.3 years (range 1.5–3.1). NLP-PAC, after adaptation, and human abstractor identified 74 (25%) and 72 subjects (24%) respectively, with 66 subjects identified by both approaches. Sensitivity, specificity, positive predictive value and negative predictive value for NLP algorithm in predicting asthma status were 92%, 96%, 89%, and 97%, respectively. The known risk factors for asthma identified by NLP (e.g., smoking history) were similar to the ones identified by manual chart review.

[¶]These authors both are responsible for correspondence: Young J. Juhn, MD, MPH, Professor of Pediatrics, Department of Pediatric and Adolescent Medicine, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905, Juhn.young@mayo.edu, Phone: 507-538-1642, Fax: 507-284-9744; Hongfang Liu, PhD, Professor of Biomedical Informatics, Division of Biomedical Statistics and Informatics, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905, Liu.hongfang@mayo.edu, Phone: 507-293-0057, Fax: 507-284-1516.

[*]These authors contributed equally

Conclusions—Successful implementation of NLP-PAC for asthma ascertainment in two different practice settings demonstrates the feasibility of automated asthma ascertainment leveraging EHR data with a potential to enable large scale, multi-site asthma studies to improve asthma care and research.

Keywords

informatics; retrospective study; electronic health records; validation; natural language processing; algorithm adaptability; asthma ascertainment; epidemiology

Introduction

Asthma is the most common chronic illness in childhood. [1, 2] Asthma poses significantly increased risks of serious or common microbial infections in addition to its own morbidities.[3–11] However, a significant delay in asthma diagnosis frequently occurs, delaying timely access to preventive and therapeutic interventions for asthma.[12–15]

Despite the availability of electronic health records (EHRs), several major barriers remain as impediments to research and improved care for asthma leveraging EHR data. Structured data (e.g., ICD-9 codes) lack the accuracy to effectively identify and manage asthmatic children in real time (e.g., sensitivity 31% with the predetermined asthma criteria as reference).[16] Manual chart reviews for asthma ascertainment are labor-intensive, and thus not feasible to apply for large scale studies or clinical practice as a population management tool, although it has been widely used for epidemiological studies.[17–21] To address these barriers, we, as a multidisciplinary research team, developed and validated a natural language processing (NLP) algorithm, NLP-PAC, that automatically ascertains asthma status using EHR data based on our predetermined asthma criteria (PAC), allowing early identification and treatment for asthmatic children.[16, 22] Kappa index and agreement for asthma status between NLP-PAC and manual chart review were 0.85 and 0.95, respectively, suggesting excellent performance.[23] At present, the NLP-PAC has not been applied (i.e., adapted) to other institutions outside Mayo Clinic (i.e., unknown external validity).

Herein, we aim to investigate the external validity of NLP-PAC, originally developed at Mayo Clinic, using a birth cohort at Sanford Children Hospital (SCH) and report our findings.

Methods

Study setting

SCH is one of the major health care systems in South Dakota serving the counties of Minnehaha and Lincoln. Any child born at SCH is usually followed over time by pediatric and family medicine providers who are affiliated with this institution. All routine and acute care is documented in the EPIC EHR system that was first implemented in 2011. Unique identifiers are assigned to each patient and sub-identifiers are created for each subsequent visit. Also, each visit mandates the provider to record a visit-related diagnosis. Having these identifiers proved useful for data collection in this study.

Study design

This is a retrospective cohort study based on a sample of the 2011–2012 SCH Birth cohort (n=595). The main aim of this study was to assess external validity of NLP-PAC, which was developed and validated at Mayo Clinic Rochester, by adapting the NLP-PAC for the SCH HER data. Manual chart review was performed on the cohort for asthma ascertainment based on PAC. We then used half of the cohort as a training cohort (n=298) to adapt NLP-PAC for SCH and the other half served as a blind test cohort to evaluate the adapted NLP-PAC algorithm. Specifically, using the training cohort, we performed an error-analysis for false positives (i.e., NLP indicates yes for asthma, but abstractor (EK) indicates no) and false negatives (i.e., vice versa) to revise and refine NLP-PAC through a reiterative process. Then, we ran the adapted NLP-PAC on the blind test cohort to assess criterion validity and construct validity of the adapted NLP-PAC. Criterion validity was assessed by determining concordance of asthma status by PAC between the adapted NLP-PAC and manual chart review. Construct validity of the adapted NLP-PAC was assessed by determining the association between asthma status ascertained by NLP algorithms and the known risk factors for asthma.

Study subjects

The study cohort at Mayo Clinic that was utilized for the development and internal validation of the original NLP algorithm was previously described in detail in our original study.[16, 23] Briefly, for the test cohort of Mayo Clinic, we utilized a random sample of 500 subjects from the Olmsted County Birth Cohort, 1997–2007, who were born after Mayo EHR implementation and have had primary care at Mayo Clinic. In the original study, we selected study subjects from a birth cohort as a sampling frame to minimize sampling bias. Similarly, to assess external validity of NLP-PAC, in this study, we enrolled a birth cohort of 1,549 children who were born between 11/1/2011 and 10/31/2012 and had at least two well-child exams at SCH between 11/1/2011 and 10/31/2013. We chose this birth cohort because 1) the birth cohort has comprehensive medical records during early childhood which is important to apply asthma criteria to identify children with asthma (i.e. not by self-report), 2) SCH started their EHR system in 2011, and 3) having primary care at SCH, ideally for longer follow up is necessary to identify children with asthma by capturing all potential asthma-related visits from SCH's EHR.

Predetermined Asthma Criteria (PAC)

Drs. John Yunginger and Charles Reed, renowned researchers and clinicians for asthma at Mayo Clinic, developed and validated the original Predetermined Asthma Criteria (PAC) for retrospective studies among children and adults based on chart review (Table 1).[24] Although the PAC was not designed to replace or prompt diagnosing asthma in clinical practice, the PAC includes recurrent wheezing symptoms along with other respiratory symptoms (e.g., night cough, difficulty in breathing) and airway hyperresponsiveness suggested as key symptom indicators for considering a diagnosis of asthma by the National Asthma Education and Prevention Program Expert Panel Report-3.[25] To our knowledge, these criteria are the only existing predetermined criteria for asthma that determines asthma status and the index date of incident asthma retrospectively based on medical records. As defined

by PAC, most cases of probable asthma (85%) became definite asthma over time, so we included both definite and probable asthma for the present study.[24, 26] PAC was found to have high reliability and extensive epidemiologic work for asthma has used PAC showing the excellent construct validity in identifying known risk factors for asthma and asthma-related adverse outcomes (e.g., microbial infections).[26–36]

Asthma ascertainment by abstractor and NLP

Asthma status was determined using PAC by an abstractor (EK) using EHR data available from birth to the last follow-up date. The development of NLP-PAC was previously described in detail.[16, 22] Briefly, we first created a rule-based NLP algorithm for PAC delineated in Table 1. To determine asthma status by the NLP algorithms, we used a two-step process, including a text processing component for medical records (finding asthma-related concepts in text that match the specified criteria) and a patient classification component (deciding the asthma status of a patient based on the available evidence from the text processing step). NLP-PAC is not a simple keyword search but a combination of rules utilizing assertion status (e.g., negation, possible, associated with patient), section constraint (e.g., diagnosis), temporal association constraint (e.g., wheezing and coughing should occur at the same time), and note types (e.g., exclude notes from unrelated practice settings). For example, if NLP-PAC encountered a sentence that stated “no rales or wheezing,” it marked that the concept “wheezing” was found but recognized that the patient did not have wheezing (i.e., negated rule); a physician diagnosis of asthma was extracted specifically from the diagnosis section and asthma diagnosis from the family history section was not considered because it is not patient-specific (i.e., section constraint rule).

Adaptation of NLP-PAC for SCH

SCH uses an EPIC EHR system which generates EHR data with different text format, section definitions, and note types in practice settings compared to Mayo Clinic. However, asthma-related concepts in medical records required for PAC are similar between the two institutions that allow an NLP algorithm to be adaptable to SCH once we adjust those variations. To adapt NLP-PAC to SCH, we first pre-processed SCH text format (e.g., sentence format) to be applicable for NLP-PAC and determined which sections and note types needed to be included or excluded through discussion with Mayo and SCH clinicians. Additionally, we refined the assertion identification to correctly handle some different negation patterns used in SCH, which will in turn enhance the NLP text processing performance. For example, for “wheezing, none recently,” where wheezing is negated by negation keywords “none recently,” this negation pattern was not defined in the original NLP-PAC but added when applied to SCH.

Other variables—The known risk factors for asthma such as a family history of asthma, a history of allergic rhinitis or eczema, maternal smoking during pregnancy, passive smoking after birth, and breastfeeding as well as tympanostomy tube insertion and Streptococcus pyogenes infection during study follow-up period, which are known to be associated with asthma,[37, 38] were collected by abstractor (EK).

Statistical analysis

Performance of the adapted NLP-PAC as of the last follow-up date was assessed for criterion validity (manual chart review as a gold standard) with regard to kappa index, agreement rate, sensitivity, specificity, positive predictive value, and negative predictive value. Logistic regression models were utilized to assess the association between known risk factors for asthma and asthma status based on each ascertainment criteria (construct validity). All analyses were performed using JMP statistical software package (Ver 10; SAS Institute, Inc, Cary, NC).

Results

Characteristics of study subjects

Among eligible 297 study subjects (i.e., test cohort), 160 (53%) were males, 268 (90%) White, and the median age at last follow-up date was 2.3 years (IQR 2.0–2.6 and range 1.5–3.1). 37 children (12%) had a family history of asthma, and 12 (4%) and 69 (23%) had allergic rhinitis and eczema, respectively. 24 subjects (8%) had physician diagnosis of asthma during the study period.

Concordance in asthma status between the adapted NLP-PAC and manual chart review (criterion validity)

The results are summarized in Table 2. The NLP algorithm identified 74 subjects (25%) meeting PAC while manual chart review identified 72 (24%) subjects with 66 identified by both approaches. Kappa index and agreement for asthma status between NLP algorithm and manual chart review were 0.87 and 0.95, respectively suggesting excellent agreement. Sensitivity, specificity, positive predictive value, and negative predictive value for the algorithm against asthma status ascertained by manual chart review were 92%, 96%, 89%, and 97%, respectively. Most error of the adapted NLP-PAC was due to incorrect negation of asthma-related concepts.

Association of asthma status determined by the adapted NLP-PAC and manual chart review in relation to the known risk factors and outcomes for asthma (construct validity)

The known risk factors for asthma identified by NLP were almost the same as the ones identified by the abstractor (Table 3). Children with asthma determined by NLP compared to those without had higher odds of having a family history of asthma, a history of allergic rhinitis and eczema, maternal smoking during pregnancy, and passive smoking after birth ($p<.05$ in each). Asthma status by manual chart review also showed similar results with regard to the association with known risk factors for asthma, except eczema with marginal significance (OR (95%CI), p-value: 1.5 (0.8–2.7), p=.17). The association of asthma status with no history of breastfeeding approached to significance (OR (95%CI), p-value: 0.5 (0.2–1.1), p=.12 for NLP, and 0.5 (0.2–1.0), p=.07 for manual chart review). Children with NLP-identified asthma had more than three times higher odds of *Streptococcus pyogenes* upper respiratory infection and tympanostomy tube insertion rate, which is found to be similar among those with manual chart review-defined asthma status.

Discussion

We demonstrated automated asthma ascertainment based on EHR data leveraging NLP is feasible across multiple sites. Previous research has been limited by inconsistent asthma definitions and ascertainment processes, which impact biological precision and obscure the true biological heterogeneity of asthma status and prognosis. The use of specific asthma ascertainment criteria and the associated NLP algorithm in multi-site studies will reduce the heterogeneity of asthma ascertainment and improve biological precision in detecting true epidemiologic findings.

Difficulty of accurately determining asthma status is a significant impediment to population-based asthma research, as suggested by the 2014 NIH-led Joint Workshop for Birth Cohorts.[39] Inconsistent results have been reported in genome-wide association studies (GWAS),[40–42] clinical trials,[43, 44] and studies addressing heterogeneity of asthma.[45–49] A recent National Heart Lung Blood Institute workshop discussed the core and supplementary predictor and outcome variables for asthma research,[50] but left asthma criteria, ascertainment processes, and choice of sampling frames *undefined*, thus permitting inconsistent practices in ascertaining asthma and selecting study samples to continue. To address this challenge, we developed and validated an NLP algorithm for asthma status at Mayo Clinic,[16, 22] and demonstrated external validity (generalizability) at SCH in this study. Our study demonstrated the NLP algorithm can be adapted in a different care setting with comparable performance which will enable us to standardize asthma definitions.

Criteria-based asthma ascertainment is helpful for identifying patients with undiagnosed asthma despite recurrent asthma symptoms. However, with a large-scale study, manual chart review for applying asthma criteria is time-consuming, inefficient, and sometimes inaccurate, especially for patients with a large volume of medical records. Alternatively, billing codes (e.g. ICD 9 or 10) with poor sensitivity have been used to identify people with asthma, resulting in missing those with undiagnosed asthma, especially children less than 3 years old for whom clinicians may be reluctant to label such children with asthma.[16, 27] Among 72 subjects who met PAC in this study, only 24 had a physician diagnosis of asthma (33%). Given the younger age of the cohort (median age 2.3 years (IQR 2.0–2.6 and range 1.5–3.1)), this may not be surprising, but this is one of the reasons why children this age are often not included in large-scale asthma studies.[51, 52] Criteria-based NLP algorithm for asthma will help identify children with recurrent asthma symptoms regardless of age on a population level. Asthma status defined by NLP algorithms at Mayo Clinic and SCH showed significant association with known risk factors for asthma such as family history of asthma or allergic rhinitis. While PAC is a useful clinical decision support tool for a clinician in identifying children with potential cases of asthma, it does not replace clinician's clinical judgment for asthma diagnosis in which clinicians might disagree with case definition of asthma by NLP for certain patients.

We demonstrated that NLP-PAC, originally developed at Mayo Clinic, can be adapted to another health care setting and that the adapted NLP-PAC is effective (compared to manual chart review) in ascertaining asthma status. The major effort for this project was for manual chart review. After pre-processing medical records into a common format, the effort taken to

adapt NLP-PAC for SCH was to apply the algorithm on a cohort and then perform error-analysis to identify missing patterns. Roughly, apart from the project leadership, 1 medical informatics staff (0.1 FTE) at Mayo Clinic and 1 IT specialist (0.05 FTE) at Sanford Children's Hospital were involved with the process adapting the original NLP algorithm for SCH for about six months. This is much less effort than developing and validating the original system (approximately 1 year with 0.3 FTE of the same medical informatician). There was no significant barrier to apply the existing NLP algorithm developed by Mayo into SCH although there are variations of text format, section definitions, and note types between two institutions, resulting in some discrepancies between NLP and abstractor. Those variations were able to be handled appropriately through adjusting text format, and reconciling section and note type differences by mutual communication between the two institutions. It will be worth it to make the adapting process to other institutions more efficient given potentials and benefits of using the once developed or adapted NLP algorithm regarding asthma research and practice with the least cost and effort compared to manual chart review. Currently, we are investigating the possible mechanisms to implement this algorithm at Mayo Clinic and other institutions.

NLP used in health care systems has evolved to aid clinical decision making of health care providers by providing easily accessible health-related information.[53] However, NLP use in respiratory care is limited to extracting an individual clinical finding or risk factor (e.g., pulmonary function test, smoking status),[54, 55] not a patient classification by applying complex criteria (i.e., deciding asthma status based on available evidence) beyond a text processing component (i.e., finding evidence text in EHRs to match specified criteria). To our knowledge, our validated NLP algorithm is the first and only algorithm to identify asthma status by applying asthma criteria.

The main strength of this study is a demonstration of the NLP-PAC algorithm adaptability across institutions. Another strength is using a birth cohort with medical records of early childhood available, which makes it comprehensive to apply criteria to EHR for asthma ascertainment. Challenges of this study include identifying and consolidating EHR data variations between two institutions in a cohesive manner for an NLP algorithm. Also, sharing EHR documents between two health care enterprises required some technical and administrative effort, which may also be required applying this algorithm to other health care institutions. The portability issue between two different institutions using the same EHR system (e.g., EPIC) might not require the same level of validation process as that for two different EHR systems. However, it would be still necessary to perform some validation including chart annotation by abstractor followed by error-analysis for false positives and false negatives whenever this algorithm is implemented in a new practice setting. But once the algorithms is validated at a new practice setting, it will be much more efficient and cost-effective to run the algorithm to identify the PAC-met patients in almost real-time, compared to manual chart review as our study finding suggested. Lastly, while the PAC has been used for adult studies successfully and there is potential for using NLP-PAC in adults, this NLP-PAC algorithm has not yet been applied to an adult cohort. So, different approaches may be warranted to be developed for studies including adults who may have some portion of non-EHR charts (i.e., paper charts).

For different prevalence of asthma by NLP between the two health care systems (24% at Sanford vs. 31% at Mayo), potential systemic differences between the two sites could be speculated in addition to different age distribution of different cohorts (median age 2.3 years (range: 1.5–3.1) for SCH vs. 11.5 years (range: 4.7–17.9) for Mayo Clinic) as the availability of EHR documents and their contents which are resources for applying the PAC are dependent upon the patient age. Although it is hard to define the degree of source of heterogeneity in difference in prevalence of asthma in two sites, it could be due to actual difference in asthma prevalence, practice (clinician) difference (e.g., pattern of using the term of wheezing, dictating vs. typing) or EHR system difference (e.g. sectionizing).

In conclusion, successful implementation of automated asthma ascertainment based on EHR leveraging NLP into two health care settings was feasible by establishing viable and cohesive multidisciplinary teamwork. This NLP algorithm for asthma research and care has strong potential for automated chart review which enables large-scale population studies, timely asthma diagnosis, a reduction in the delay of asthma diagnosis, has possibility of real-time asthma surveillance in clinical practice, and thus, will improve overall asthma care as a population management tool.

Acknowledgments

We thank IT staff (Denise Schoolmeester, Carmen Petersonat) at SCH for their technical support and Mrs. Kelly Okeson for her administrative assistance. This work was supported by NIH-funded R01 grant (R01 HL126667), R21 grant (R21AI116839-01), and T. Denny Sanford Pediatric Collaborative Research Fund. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH.

Abbreviations

NLP	Natural Language Processing
EHR	Electronic Health Record
SCH	Sanford Children’s Hospital
PAC	Predetermined Asthma Criteria

References

1. Lethbridge-Cejku M, Vickerie J. Summary of Health Statistics for US Adults: National Health Interview Survey, 2003. National Center for Health Statistics. 2005; 10(225):2005.
2. Anonymous. Forecasted state-specific estimates of self-reported asthma prevalence–United States, 1998. MMWR – Morbidity & Mortality Weekly Report. 1998; 47:1022–5. [PubMed: 9853939]
3. Talbot T, Hartert TV, Arbogast PG, Mitchel E, Schaffner K, Craig AS, Griffin MR. Asthma as a Risk Factor for Invasive Pneumococcal Disease. New England Journal of Medicine. 2005; 352:2082–90. [PubMed: 15901861]
4. Juhn YJ, Kita H, Yawn BP, Boyce TG, Yoo KH, McGree ME, et al. Increased risk of serious pneumococcal disease in patients with asthma. Journal of Allergy and Clinical Immunology. 2008; 122:719–23. [PubMed: 18790525]
5. Capili CHA, Rigelman-Hedberg N, Fink L, Boyce T, Juhn Y. Increased Risk of Pertussis in Patients with Asthma. Journal of Allergy & Clinical Immunology. 2012; 129:957–63. [PubMed: 22206778]

6. Jain S, Kamimoto L, Bramley AM, Schmitz AM, Benoit SR, Louie J, et al. Hospitalized patients with 2009 H1N1 influenza in the United States, April–June 2009. N Engl J Med. 2009; 361:1935–44. [PubMed: 19815859]
7. Webb SA, Pettila V, Seppelt I, Bellomo R, Bailey M, Cooper DJ, et al. Critical care services and 2009 H1N1 influenza in Australia and New Zealand. N Engl J Med. 2009; 361:1925–34. [PubMed: 19815860]
8. Kloepfer JPO KM, Lee W, Pappas TE, Liu G, Vrtis RF, Evans MD, Gangnon RE, Gern JE. Increased H1N1 Infection Rate in Asthmatic Children. The Journal of Allergy and Clinical Immunology. 2011; 127 AB147 #555.
9. Frey DM, Li X, Weaver AL, Jacobson RM, Poland GA, Juhn YJ. Increased Risk of Group A Streptococcal Upper Respiratory Infections in Pediatric Asthmatics. The Journal of Allergy and Clinical Immunology. 2008; 121:S23.
10. Bjur KALR, Fenta Y, Yoo KH, Li X, Jacobson RM, Juhn YJ. Assessment of the Association Between Atopic Conditions and Tympanostomy Tube Placement in Children. Allergy and Asthma Proceedings. 2012 In Press.
11. Kim, B., Mehra, S., Yawn, B., Terrell, R., Lahr, B., Juhn, YJ. Asthma and risk of herpes zoster in children: a population-based case-control study. The 2012 Annual Pediatric Academic Society Meeting; Boston, MA. 2012;
12. Molis WE, Bagniewski S, Weaver AL, Jacobson RM, Juhn YJ. Timeliness of diagnosis of asthma in children and its predictors. Allergy. 2008; 63:1529–35. [PubMed: 18925889]
13. Lynch BA, Van Norman CA, Jacobson RM, Weaver AL, Juhn YJ. Impact of delay in asthma diagnosis on health care service use. Allergy and asthma proceedings: the official journal of regional and state allergy societies. 2010; 31:e48–e52.
14. Bisgaard H, Szefler S. Prevalence of asthma-like symptoms in young children. Pediatric Pulmonology. 2007; 42:723–8. [PubMed: 17598172]
15. Davoodi POD, Havstad S, Waller J, Joseph C, Tingen M. Characteristics of Adolescents with Undiagnosed Asthma in Rural Counties in Georgia. AAAAI Abstract. 2014
16. Wu ST, Sohn S, Ravikumar KE, Wagholikar K, Jonnalagadda SR, Liu H, et al. Automated chart review for asthma cohort identification using natural language processing: an exploratory study. Annals of allergy, asthma & immunology: official publication of the American College of Allergy, Asthma, & Immunology. 2013; 111:364–9.
17. Juhn YJ, Kita H, Yawn BP, Boyce TG, Yoo KH, McGree ME, et al. Increased risk of serious pneumococcal disease in patients with asthma. The Journal of allergy and clinical immunology. 2008; 122:719–23. [PubMed: 18790525]
18. Capili CR, Hettinger A, Rigelman-Hedberg N, Fink L, Boyce T, Lahr B, et al. Increased risk of pertussis in patients with asthma. The Journal of allergy and clinical immunology. 2012; 129:957–63. [PubMed: 22206778]
19. Bjur KA, Lynch RL, Fenta YA, Yoo KH, Jacobson RM, Li X, et al. Assessment of the association between atopic conditions and tympanostomy tube placement in children. Allergy and asthma proceedings: the official journal of regional and state allergy societies. 2012; 33:289–96.
20. Wi CI, Park MA, Juhn YJ. Development and initial testing of Asthma Predictive Index for a retrospective study: an exploratory study. The Journal of asthma: official journal of the Association for the Care of Asthma. 2015; 52:183–90. [PubMed: 25158051]
21. Wi CI, Kim BS, Mehra S, Yawn BP, Park MA, Juhn YJ. Risk of herpes zoster in children with asthma. Allergy and asthma proceedings: the official journal of regional and state allergy societies. 2015; 36:372–8.
22. Wu ST, Juhn YJ, Sohn S, Liu H. Patient-level temporal aggregation for text-based asthma status ascertainment. Journal of the American Medical Informatics Association: JAMIA. 2014; 21:876–84. [PubMed: 24833775]
23. Sohn, Wi C., Liu, S., Ryu, H., Park, E., Juhn, MY. 2016 American Academy of Allergy, Asthma & Immunology. Los Angeles, CA: 2016. Automated Chart Review for Asthma Ascertainment: An Innovative Approach for Asthma Care and Research in the Era of Electronic Medical Record. Abstract

24. Yunginger J, Reed CE, O'Connell EJ, Melton J, O'Fallon WM, Silverstein MD. A Community-based Study of the Epidemiology of Asthma: Incidence Rates, 1964–1983. Am Rev Respir Dis. 1992; 146:888–94. [PubMed: 1416415]
25. Expert Panel Report 3 (EPR-3). Guidelines for the Diagnosis and Management of Asthma-Summary Report 2007. The Journal of allergy and clinical immunology. 2007; 120:S94–138. [PubMed: 17983880]
26. Juhn YJ, Kita H, Lee LA, Swanson RJ, Smith R, Bagniewski SM, et al. Childhood asthma and measles vaccine response. Annals of allergy, asthma & immunology: official publication of the American College of Allergy, Asthma, & Immunology. 2006; 97:469–76.
27. Yunginger JW, Reed CE, O'Connell EJ, Melton LJ 3rd, O'Fallon WM, Silverstein MD. A community-based study of the epidemiology of asthma. Incidence rates, 1964–1983 The American review of respiratory disease. 1992; 146:888–94. [PubMed: 1416415]
28. Beard CM, Yunginger JW, Reed CE, O'Connell EJ, Silverstein MD. Interobserver variability in medical record review: an epidemiological study of asthma. Journal of clinical epidemiology. 1992; 45:1013–20. [PubMed: 1432015]
29. Hunt LW Jr, Silverstein MD, Reed CE, O'Connell EJ, O'Fallon WM, Yunginger JW. Accuracy of the death certificate in a population-based study of asthmatic patients. JAMA : the journal of the American Medical Association. 1993; 269:1947–52. [PubMed: 8464126]
30. Silverstein MD, Reed CE, O'Connell EJ, Melton LJ 3rd, O'Fallon WM, Yunginger JW. Long-term survival of a cohort of community residents with asthma. The New England journal of medicine. 1994; 331:1537–41. [PubMed: 7969322]
31. Bauer BA, Reed CE, Yunginger JW, Wollan PC, Silverstein MD. Incidence and outcomes of asthma in the elderly. A population-based study in Rochester, Minnesota. Chest. 1997; 111:303–10. [PubMed: 9041973]
32. Silverstein MD, Yunginger JW, Reed CE, Petterson T, Zimmerman D, Li JT, et al. Attained adult height after childhood asthma: effect of glucocorticoid therapy. The Journal of allergy and clinical immunology. 1997; 99:466–74. [PubMed: 9111490]
33. Juhn YJ, Qin R, Urm S, Katusic S, Vargas-Chanes D. The influence of neighborhood environment on the incidence of childhood asthma: a propensity score approach. The Journal of allergy and clinical immunology. 2010; 125:838–43. e2. [PubMed: 20236695]
34. Juhn YJ, Sauver JS, Katusic S, Vargas D, Weaver A, Yunginger J. The influence of neighborhood environment on the incidence of childhood asthma: a multilevel approach. Social science & medicine. 2005; 60:2453–64. [PubMed: 15814171]
35. Juhn YJ, Weaver A, Katusic S, Yunginger J. Mode of delivery at birth and development of asthma: A population-based cohort study. Journal of Allergy and Clinical Immunology. 2005; 116:510–6. [PubMed: 16159617]
36. Yawn BP, Yunginger JW, Wollan PC, Reed CE, Silverstein MD, Harris AG. Allergic rhinitis in Rochester, Minnesota residents with asthma: frequency and impact on health care charges. The Journal of allergy and clinical immunology. 1999; 103:54–9. [PubMed: 9893185]
37. Juhn YJ, Wi CI. What does tympanostomy tube placement in children teach us about the association between atopic conditions and otitis media? Current allergy and asthma reports. 2014; 14:447. [PubMed: 24816652]
38. Frey D, Jacobson R, Poland G, Li X, Juhn Y. Assessment of the association between pediatric asthma and Streptococcus pyogenes upper respiratory infection. Allergy Asthma Proc. 2009; 30:540–5. [PubMed: 19674512]
39. Bousquet J, Gern JE, Martinez FD, Anto JM, Johnson CC, Holt PG, et al. Birth cohorts in asthma and allergic diseases: report of a NIAID/NHLBI/MeDALL joint workshop. The Journal of allergy and clinical immunology. 2014; 133:1535–46. [PubMed: 24636091]
40. Xingnan L, Timothy DH, Siqun LZ, Tmirah H, Stephen PP, Deborah AM, et al. Genome-wide association study of asthma identifies RAD50-IL13 and HLA-DR/DQ regions. The Journal of allergy and clinical immunology. 2010; 125:328–35.e11. [PubMed: 20159242]
41. Ferreira MA, Matheson MC, Duffy DL, Marks GB, Hui J, Le Souef P, et al. Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. Lancet. 2011; 378:1006–14. [PubMed: 21907864]

42. Deborah AM. Genetics of asthma and allergy: What have we learned? The Journal of allergy and clinical immunology. 2010; 126:439–46. [PubMed: 20816180]
43. Ducharme FM, Lemire C, Noya FJD, Davis GM, Alos N, Leblond H, et al. Preemptive Use of High-Dose Fluticasone for Virus-Induced Wheezing in Young Children. N Engl J Med. 2009; 360:339–53. [PubMed: 19164187]
44. Panickar J, Lakhanpaul M, Lambert PC, Kenia P, Stephenson T, Smyth A, et al. Oral Prednisolone for Preschool Children with Acute Virus-Induced Wheezing. N Engl J Med. 2009; 360:329–38. [PubMed: 19164186]
45. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, et al. Cluster Analysis and Clinical Asthma Phenotypes. Am J Respir Crit Care Med. 2008; 178:218–24. [PubMed: 18480428]
46. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. Am J Respir Crit Care Med. 2010; 181:315–23. [PubMed: 19892860]
47. Fitzpatrick AM, Teague WG, Meyers DA, Peters SP, Li X, Li H, et al. Heterogeneity of severe asthma in childhood: Confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. The Journal of Allergy and Clinical Immunology. 2011; 127:382–9.e13. [PubMed: 21195471]
48. Lazic N, Roberts G, Custovic A, Belgrave D, Bishop CM, Winn J, et al. Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. Allergy. 2013; 68:764–70. [PubMed: 23621120]
49. Pratt GC, Parson K, Shinoda N, Lindgren P, Dunlap S, Yawn B, et al. Quantifying traffic exposure. Journal of exposure science & environmental epidemiology. 2014; 24:290–6. [PubMed: 24045427]
50. Busse WW, Morgan WJ, Taggart V, Togias A. Asthma outcomes workshop: Overview. The Journal of allergy and clinical immunology. 2012; 129:S1–S8. [PubMed: 22386504]
51. The Childhood Asthma Management Program (CAMP): design, rationale, and methods. Childhood Asthma Management Program Research Group. Controlled clinical trials. 1999; 20:91–120. [PubMed: 10027502]
52. Liu AH, Gilsenan AW, Stanford RH, Lincourt W, Ziemiecki R, Ortega H. Status of asthma control in pediatric primary care: results from the pediatric Asthma Control Characteristics and Prevalence Survey Study (ACCESS). The Journal of pediatrics. 2010; 157:276–81. e3. [PubMed: 20472251]
53. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? Journal of Biomedical Informatics. 2009; 42:760–72. [PubMed: 19683066]
54. Sauer BC, Jones BE, Globe G, Leng J, Lu CC, He T, et al. Performance of a Natural Language Processing (NLP) Tool to Extract Pulmonary Function Test (PFT) Reports from Structured and Semistructured Veteran Affairs (VA) Data. EGEMS (Wash DC). 2016; 4:1217. [PubMed: 27376095]
55. Meystre SM, Deshmukh VG, Mitchell J. A clinical use case to evaluate the i2b2 Hive: predicting asthma exacerbations. AMIA Annu Symp Proc. 2009; 2009:442–6. [PubMed: 20351896]

Highlights

1. What is already known about this topic?

A natural language processing (NLP) algorithm for asthma ascertainment based on predetermined asthma criteria (PAC) leveraging EHR data, NLP-PAC, provides an opportunity for early identification and treatment for children with asthma in one institution.

2. What does this article add to our knowledge?

We successfully adapted the algorithm to ascertain asthma in a different care setting with much less effort and demonstrated external validity and adaptability.

3. How does this study impact current management guidelines?

Automated asthma ascertainment based on EHR will enable large scale, multi-site asthma studies to improve asthma care and research by minimizing methodological heterogeneity stemming from different asthma ascertainment processes.

Table 1

Predetermined Asthma Criteria

Patients were considered to have *definite* asthma if a physician had made a diagnosis of asthma and/or if each of the following three conditions were present, and they were considered to have *probable* asthma if only the first two conditions were present:

1 History of cough with wheezing, and/or dyspnea, OR history of cough and/or dyspnea plus wheezing on examination,

2 Substantial variability in symptoms from time to time or periods of weeks or more when symptoms were absent, and

3 Two or more of the following:

- Sleep disturbance by nocturnal cough and wheeze
- Nonsmoker (14 years or older)
- Nasal polyps
- Blood eosinophilia higher than 300/uL
- Positive wheal and flare skin tests OR elevated serum IgE
- History of hay fever or infantile eczema OR cough, dyspnea, and wheezing regularly on exposure to an antigen
- Pulmonary function tests showing one FEV_1 or FVC less than 70% predicted and another with at least 20% improvement to an FEV_1 of higher than70% predicted OR methacholine challenge test showing 20% or greater decrease in FEV_1
- Favorable clinical response to bronchodilator

Patients were excluded from our previous study if any of these conditions were present:

- Pulmonary function tests that showed FEV_1 to be consistently below 50% predicted or diminished diffusion capacity
- Tracheobronchial foreign body at or about the incidence date
- Hypogammaglobulinemia (IgG less than 2.0 mg/mL) or other immunodeficiency disorder
- Wheezing occurring only in response to anesthesia or medications
- Bullous emphysema or pulmonary fibrosis on chest radiograph
- PiZZ $alpha_1$-antitrypsin
- Cystic fibrosis
- Other major chest disease such as juvenile kyphoscoliosis or bronchiectasis FVC forced vital capacity; FEV1, forced expiratory volume in 1 sec.

Table 2

Agreement for asthma status between NLP algorithm and manual chart review (gold standard)

Unweighted Cohen's kappa	Overall agreement	Sensitivity	Specificity	Positive predictive value	Negative predictive value
0.87	0.95	66/72 (92%)	217/225 (96%)	66/74 (89%)	217/223 (97%)

Table 3

Associations of asthma status determined by NLP with known risk factors for asthma (construct validity)

	By NLP				By manual chart review			
	No asthma	**Asthma**	**OR (95% CI)**	**p-value**	**No asthma**	**Asthma**	**OR (95% CI)**	**p-value**
Age,[a] years, median (range)	2.3(1.5,3.0)	2.3(1.7, 3.1)	1.4(0.6, 6.1)	.32	2.3(1.5,3.0)	2.3(1.7, 3.1)	1.5(0.7, 3.3)	.23
Male	115(51%)	45(60%)	1.4(0.8, 2.4)	.16	119(52%)	41(56%)	1.1(0.6, 2.0)	.54
White	201(90%)	67(90%)	0.9(0.3, 2.3)	1.0	202(90%)	66(91%)	1.1(0.4, 2.9)	1.0
Allergic rhinitis	4(1%)	8(10%)	6.6(1.9, 22.7)	.002	4(1%)	8(11%)	6.9(2.0, 23.6)	<.001
Eczema	45(20%)	24(32%)	1.8(1.0, 3.4)	.03	48(21%)	21(29%)	1.5(0.8, 2.7)	.17
Family history of asthma	18(8%)	19(25%)	3.9(1.9, 8.0)	<.001	19(8%)	18(25%)	3.6(1.7, 7.3)	<.001
Smoking during pregnancy[b]	6(2%)	3(4%)	1.5(0.3, 6.2)	.69	6(2%)	3(4%)	1.5(0.3, 6.5)	.51
Passive smoking after birth[c]	20(8%)	11(14%)	1.7(0.8, 3.8)	.18	20(8%)	11(15%)	1.8(0.8, 4.0)	.12
Breastfeeding ever	195(87%)	59(79%)	0.5(0.2, 1.1)	.12	197(87%)	57(79%)	0.5 (0.2, 1.0)	.07
Tympanostomy Tube insertion ever	18(8%)	16(21%)	3.1(1.5, 6.5)	.001	17(7%)	17(23%)	3.7(1.8, 7.8)	<.001
Streptococcus pyogenes infection ever	19(8%)	18(24%)	3.4(1.6, 7.0)	<.001	22(9%)	15(20%)	2.4(1.1, 4.9)	.01

[a] Age at the last follow-up date;

[b] maternal smoking status during pregnancy;

[c] Household smoking exposure status during infant