Original Article

Natural Language Processing for Asthma Ascertainment in Different Practice Settings

Chung-Il Wi, MD[a,b,*], Sunghwan Sohn, PhD[c,*], Mir Ali, MD[d], Elizabeth Krusemark, AAS[a,b], Euijung Ryu, PhD[c], Hongfang Liu, PhD[c], and Young J. Juhn, MD, MPH[a,b] *Rochester, Minn; and Sioux Falls, SD*

What is already known about this topic? A natural language processing (NLP) algorithm for asthma ascertainment based on predetermined asthma criteria (PAC) leveraging electronic health record data, NLP-PAC, provides an opportunity for early identification and treatment for children with asthma in one institution.

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.

How does this study impact current management guidelines? Automated asthma ascertainment based on electronic health record will enable large-scale, multisite asthma studies to improve asthma care and research by minimizing methodological heterogeneity stemming from different asthma ascertainment processes.

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 at Mayo Clinic.

OBJECTIVE: To adapt NLP-PAC in a different health care setting, Sanford Children Hospital, by assessing its external validity.

METHODS: The study was designed as a retrospective cohort study that used a random sample of 2011-2012 Sanford Birth cohort (n = 595). Manual chart review was performed on the cohort for asthma ascertainment on the basis of the 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 years). NLP-PAC, after adaptation, and the human abstractor identified 74 (25%) and 72 (24%) subjects, respectively, with 66 subjects identified by both approaches. Sensitivity, specificity, positive predictive value, and negative predictive value for the NLP algorithm in predicting asthma status were 92%, 96%, 89%, and 97%, respectively. The known risk factors for asthma identified by NLP (eg, smoking history) were similar to the ones identified by manual chart review.

CONCLUSIONS: Successful implementation of NLP-PAC for asthma ascertainment in 2 different practice settings demonstrates the feasibility of automated asthma ascertainment leveraging electronic health record data with a potential to enable large-scale, multisite asthma studies to improve asthma care and research. © 2017 American Academy of Allergy, Asthma & Immunology (J Allergy Clin Immunol Pract 2017;■:■-■)

Key words: *Informatics; Retrospective study; Electronic health records; Validation; Natural language processing; Algorithm adaptability; Asthma ascertainment; Epidemiology*

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 (eg, *International Classification of Diseases, Ninth Revision* codes)

[a]Department of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, Minn
[b]Asthma Epidemiology Research Unit, Mayo Clinic, Rochester, Minn
[c]Department of Health Sciences Research, Mayo Clinic, Rochester, Minn
[d]Department of Pediatrics, Sanford Children's Hospital, Sioux Falls, SD

This work was supported by the National Institutes of Health (NIH) (grant nos. R01 HL126667 and 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 the NIH.

Conflicts of interest: C. Wi, S. Sohn, and H. Liu have received research support from the National Institutes of Health (NIH) (grant nos. R01 HL126667 and R21AI116839-01) and T. Denny Sanford Pediatric Collaborative Research Fund. Y. J. Juhn has received research support from the NIH (grant nos. R01 HL126667 and R21AI116839-01), T. Denny Sanford Pediatric Collaborative Research Fund, and Genentech and has received consultancy fees from Genentech. The rest of the authors declare that they have no relevant conflicts of interest.

Received for publication November 9, 2016; revised March 20, 2017; accepted for publication April 12, 2017.

Available online ■■

Corresponding author: Young J. Juhn, MD, MPH, Department of Pediatric and Adolescent Medicine, Mayo Clinic, 200 1st St SW, Rochester, MN 55905 (E-mail: Juhn.young@mayo.edu); Or: Hongfang Liu, PhD, Department of Health Sciences Research, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905 (E-mail: Liu.hongfang@mayo.edu)

* These authors contributed equally to this work.

2213-2198

http://dx.doi.org/10.1016/j.jaip.2017.04.041

Abbreviations used
EHR- Electronic health record
NLP- Natural language processing
PAC- Predetermined asthma criteria
SCH- Sanford Children's Hospital

lack the accuracy to effectively identify and manage children with asthma in real time (eg, 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.[4,5,10,17,18] 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 on the basis of our predetermined asthma criteria (PAC), allowing early identification and treatment for children with asthma.[16,19] Kappa index and agreement for asthma status between NLP-PAC and manual chart review were 0.85 and 0.95, respectively, suggesting excellent performance.[20] At present, the NLP-PAC has not been applied (ie, adapted) to other institutions outside Mayo Clinic (ie, 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 first implemented in 2011. Unique identifiers are assigned to each patient and subidentifiers 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 the external validity of NLP-PAC, which was developed and validated at Mayo Clinic Rochester, by adapting the NLP-PAC for the SCH EHR 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 (ie, NLP indicates yes for asthma, but abstractor [E.K.] indicates no) and false negatives (ie, 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 used for the development and internal validation of the original NLP algorithm was previously described in detail in our original study.[16,20] Briefly, for the test cohort of Mayo Clinic, we used 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 the external validity of NLP-PAC, in this study, we enrolled a birth cohort of 1549 children who were born between November 1, 2011, and October 31, 2012, and had at least 2 well-child examinations at SCH between November 1, 2011, and October 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 (ie, not by self-report); (2) SCH started its 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

Drs John Yunginger and Charles Reed, renowned researchers and clinicians for asthma at Mayo Clinic, developed and validated the original PAC for retrospective studies among children and adults on the basis of chart review (Table I).[21] 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 (eg, night cough and 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.[22] To our knowledge, these criteria are the only existing predetermined criteria for asthma that determine asthma status and the index date of incident asthma retrospectively on the basis of 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.[21,23] 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 (eg, microbial infections).[21,23-32]

Asthma ascertainment by abstractor and NLP

Asthma status was determined using PAC by an abstractor (E.K.) using EHR data available from birth to the last follow-up date. The development of NLP-PAC was previously described in detail.[16,19] Briefly, we first created a rule-based NLP algorithm for PAC delineated in Table I. To determine asthma status by the NLP algorithms, we used a 2-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 on the basis of available evidence from the text processing step). NLP-PAC is not a simple keyword search but a combination of rules using assertion status (eg, negation, possible, and associated with patient), section constraint (eg, diagnosis), temporal association constraint (eg, wheezing and coughing should occur at the same time), and note types (eg, exclude notes from unrelated practice settings). For example, if NLP-PAC encountered a sentence that stated "no rales