

ORIGINAL REPORT

Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases

Zubair Afzal^{1*}, Marjolein Engelkes¹, Katia M. C. Verhamme¹, Hettie M. Janssens², Miriam C. J. M. Sturkenboom¹, Jan A. Kors¹ and Martijn J. Schuemie¹

¹Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

²Department of Pediatrics, Erasmus University Medical Center/Sophia Children's Hospital, Rotterdam, The Netherlands

ABSTRACT

Purpose Most electronic health record databases contain unstructured free-text narratives, which cannot be easily analyzed. Case-detection algorithms are usually created manually and often rely only on using coded information such as International Classification of Diseases version 9 codes. We applied a machine-learning approach to generate and evaluate an automated case-detection algorithm that uses both free-text and coded information to identify asthma cases.

Methods The Integrated Primary Care Information (IPCI) database was searched for potential asthma patients aged 5–18 years using a broad query on asthma-related codes, drugs, and free text. A training set of 5032 patients was created by manually annotating the potential patients as definite, probable, or doubtful asthma cases or non-asthma cases. The rule-learning program RIPPER was then used to generate algorithms to distinguish cases from non-cases. An over-sampling method was used to balance the performance of the automated algorithm to meet our study requirements. Performance of the automated algorithm was evaluated against the manually annotated set.

Results The selected algorithm yielded a positive predictive value (PPV) of 0.66, sensitivity of 0.98, and specificity of 0.95 when identifying only definite asthma cases; a PPV of 0.82, sensitivity of 0.96, and specificity of 0.90 when identifying both definite and probable asthma cases; and a PPV of 0.57, sensitivity of 0.95, and specificity of 0.67 for the scenario identifying definite, probable, and doubtful asthma cases.

Conclusions The automated algorithm shows good performance in detecting cases of asthma utilizing both free-text and coded data. This algorithm will facilitate large-scale studies of asthma in the IPCI database. Copyright © 2013 John Wiley & Sons, Ltd.

KEY WORDS—case-detection algorithms; machine learning; electronic medical records; automated case definition; pharmacoepidemiology

Received 13 August 2012; Revised 21 January 2013; Accepted 25 February 2013

INTRODUCTION

Asthma is one of the most common chronic diseases of childhood globally. The main goal of asthma treatment is to achieve and maintain clinical control of the disease. Failing to control asthma can limit daily-life activities and can be fatal. In children, asthma is usually treated and maintained with low-dose inhaled corticosteroids (ICS). If asthma is not controlled, treatment is stepped up by either adding long-acting B₂ agonists (LABAs) or a leukotriene receptor antagonist (LTRA) to low-dose ICS or increasing the dose of ICS until control is achieved.¹

Safety concerns have been raised on the long-term toxicity of ICS, the risk of mortality, and asthma exacerbations with the use of LABAs in monotherapy and the risk of neuropsychiatric events and hepatotoxicity in children treated with LTRAs.^{2–8} Randomized controlled trials on the efficacy and safety of these drugs in children are rare. In addition, the few trials conducted in children are often not designed to detect safety issues because of the limited sample size and short duration of follow-up. In general, observational studies are suited for research on drug safety because they usually have large sample size with long-term follow-up. Electronic medical records (EMRs) are valuable resources and are increasingly being used in epidemiological observational studies to detect safety issues.^{9–15}

One of the challenges of using EMRs is to determine whether and when a medical outcome of interest has occurred. When coded information such as International

*Correspondence to: Z. Afzal, Department of Medical Informatics, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands. E-mail: m.afzal@erasmusmc.nl

Classification of Diseases version 9 (ICD-9) and Logical Observation Identifiers Names and Codes (LOINC) are available, outcomes are typically identified by searching for a combination of codes in the patient record. However, the recording of these codes can be incomplete and inaccurate, or the codes themselves might be ambiguous or have the wrong granularity for the research question at hand. It is therefore recommended that the performance of this search using codes is evaluated through manual chart review, where researchers often rely on the free-text narrative in the medical record. There are also databases where the coding is simply too incomplete. For example, in the Integrated Primary Care Information (IPCI) database,¹⁶ almost 60% of the record lines are narratives and do not contain coded information. These narratives may contain important information such as patient-reported symptoms, signs, or summaries of specialists' letters. In these databases, the search for outcomes is even more labor intensive. Usually, a broad text query is defined including all possible words and codes that might be relevant, and subsequently, all narratives returned by the query are manually reviewed. With the increase in size of these databases, this practice is becoming prohibitively laborious and thus expensive.

For this reason, we used an alternative approach to identifying asthma cases, which uses the free-text narrative in an automated fashion. As suggested earlier,¹⁵ we apply a machine-learning approach to derive an automated case-detection algorithm that uses both text and coded data if available. We not only show the performance of this algorithm in terms of positive predictive value (PPV), sensitivity, and specificity but also demonstrate how sensitivity and specificity can be balanced to best meet the requirements of our study. We apply this approach to the Dutch IPCI database, but the same procedure to construct a case-detection algorithm could be used on other databases, in other languages.

METHODS

Electronic medical record database

Data in this study were taken from the IPCI database.¹⁶ The IPCI database is a longitudinal observational database of EMRs from Dutch general practitioners (GPs). The electronic records contain coded data and anonymized data on patient demographics, symptoms and diagnoses, clinical findings, referrals, laboratory findings, and hospitalization of more than 1.1 million patients. The cohort for the underlying study included children between ages 5 and 18 years that were present in the database between 1 January 2000 till 31 January 2012. A minimum

registration period of 6 months was required to guarantee sufficient medical history data.

Clinical case definition

To create a labeled training set for machine-learning methods, we used a manually defined clinical case definition. Patients were categorized into "definite asthma," "probable asthma," "doubtful asthma," or "no asthma" according to the following validation criteria.

For definite asthma patients, there was at least one entry in their medical record containing an asthma diagnosis confirmed by a specialist (pediatrician or pulmonologist). For probable asthma patients, one entry contains evidence of asthma diagnosed by the GP, and there was at least one more entry in the patient record suggestive of asthma (ICPC code, free text, or use of specific bronchodilating drugs/anti-inflammatory drugs for the indication of asthma) within the next 12 months, or there are at least two additional entries in the patient record suggestive of asthma. Use of bronchodilating drugs only did not fit these criteria. For doubtful asthma patients, there were one or more entries containing an indication or evidence of asthma, but they do not satisfy the criteria for a definite or probable asthma case. A patient is a non-asthma case if there was no indication of asthma in any entry of the patient record.

Training set for machine learning

We use machine learning methods to automatically learn case-detection algorithms on the basis of a training set of entries, that is, a set of positive and negative examples. To generate a training set for our machine-learning method, we first identified all potential asthma patients using a broad automated search on ICPC asthma disease codes, asthma drug prescriptions, and free text. The broad query is presented in the Appendix. Because of the generic asthma-related keywords used in the broad query, many of the retrieved patients were likely not to have asthma.

One medical doctor randomly reviewed the entire medical record of patients for 1 month using the clinical case definition. A total of 5032 patients were validated from 63 618 patients returned by the broad query. The doubtful patients were further reviewed by a senior medical doctor. Patients were labeled as definite asthma ($n=308$), probable asthma ($n=1133$), doubtful asthma ($n=160$), or non-asthma ($n=3431$). The entries included in the training set were the entries containing the evidence (for definite asthma cases), the entries with the GP diagnosis and the additional entries indicating asthma (for probable asthma), the entries indicating

asthma (for doubtful asthma), and a random entry from the broad query (for non-asthma cases).

A patient's medical record consists of one or more entries, where each entry pertains to a patient visit, a letter from a specialist, prescribed drugs, and so on. To make the text in the entries better fit for machine learning approaches, we removed uninformative words (so-called stop words). Although some standard Dutch stop word lists are available,¹⁷ they are not entirely suitable for the clinical text because some of the words may have importance in the clinical context; for example, "op" (English "on") could be an abbreviation of "operation." We therefore used a small stop word list, only containing "en" (English "and"), "een" ("a"), "de," and "het" (both "the").

All Anatomical Therapeutic Chemical (ATC) codes related to respiratory drugs and starting with R03 (drugs for obstructive airway diseases) were replaced by a single keyword "r03drug." To remove negated and speculative assertions, we used the Dutch assertion filter proposed in a paper by Schuemie *et al.*,¹⁵ similar to NegEx.¹⁸ Any words appearing between negation or speculation keywords and the end of sentence (demarcated by a punctuation mark) were removed from the entry. All sentences containing an alternatives keyword were completely removed.

The text in the entries was converted to lower case and split into individual words. These individual words were treated as features for our machine-learning method (bag-of-words representation). Schuemie *et al.*¹⁵ previously showed the advantage of using assertion filtering and bag-of-words representation on Dutch EMRs. For computational purposes, we reduced the number of features by chi-square feature selection.¹⁹ A *p*-value of less than 0.05 was used as a feature selection criterion. Chi-square feature selection decreased the number of features in the data set by about a factor of 10 without affecting the performance of the classifiers but greatly reducing their training time.

Automated generation of case definitions

Considering the hierarchical nature of the asthma labels (definite → probable → doubtful → non-asthma), we tackled the automated generation of case definitions as a hierarchical multi-class classification problem.^{20–22} We followed an approach in which the hierarchy is structured as a decision tree and separate classifiers are built for the nodes in the tree (Figure. 1).

We trained two machine-learning classifiers, one to separate definite cases from all other cases and the other to distinguish probable and doubtful asthma from non-asthma cases. The second classifier considered

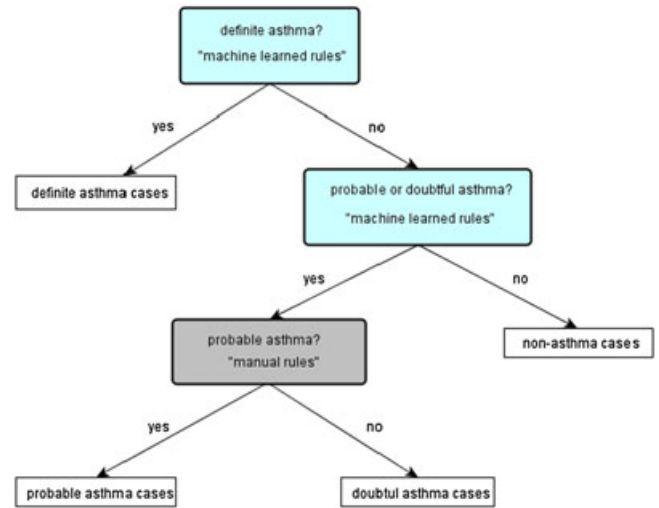


Figure 1. Hierarchical classification scheme for asthma

probable and doubtful cases as one (positive) group because the distinction between these cases is difficult to learn automatically. This distinction was made in a third classifier, which implemented two rules based on the manual case definition criteria: (1) if a patient had two positive asthma entries (according to the second classifier) within a period of 15 months and (2) if one of the entries is not a medication/drug entry, the patient was considered a probable asthma case. A medication entry only contains prescription. The training set for the first, "definite asthma" classifier consisted of the entries of the definite asthma patients as positive examples and the entries of the probable, doubtful, and non-asthma cases as negative examples. For the second, "probable, doubtful, and non-asthma" classifier, we used the entries of the probable and doubtful cases as positive examples and the entries of the non-asthma cases as negative examples. The probable or doubtful asthma patients classified as definite asthma by the first classifier were removed from the training set of the second classifier, and the definite asthma patients missed by the first classifier were added as positive examples.

To shift the balance of sensitivity and specificity, we used a method called "over-sampling." Several over-sampling methods are described in a paper by Chawla.²³ Normally, over-sampling is carried out by simply reusing the same examples multiple times, but Schuemie *et al.*¹⁵ showed that using additional entries of non-cases could lead to an increased performance. Our over-sampling was focused on increasing the specificity of the classifiers. For a non-asthma patient, all entries were manually reviewed, and no evidence of asthma was found. Although, initially, one entry was randomly selected for training the classifiers, the other entries can also be used as additional negative examples.

We created a set of all these additional entries and randomly sampled from this set to expand our training set. In total, we used 10 over-sampling percentages in the experiments. In each over-sampling run, a specified percentage of entries (of the original negative examples in the training set) from the additional entries set were added to the training set. The experiment without the over-sampling entries was considered a baseline.

Training and testing

The rule-learning algorithm RIPPER²⁴ was used on the training set to automatically generate rules for each of the asthma case definition. The RIPPER algorithm produces an ordered set of decision rules. The advantages of such machine-learning algorithms are their ability to produce output that is understandable by humans, their ease of use, and their applicability to a wide range of problems.²⁵

We used fivefold cross validation to evaluate our classifiers. Cross validation was carried out at the patient level (subject-level cross-validation²⁶); that is, the data set was randomly divided in five equally sized subsets of patients (folds). In five cross-validation runs, each time, the entries pertaining to four folds were used as a training set, and the entries of the remaining subset were used for testing. We used all entries of the patients in the test fold because in real-life situations, we do not know the labels of the entries pertaining to the patients returned by the broad query. Cross validation was used to obtain unbiased performance estimates of the classifiers, but all data was used to generate the final sets of rules.

We used PPV, sensitivity, and specificity as measures to evaluate the performance of the classifiers. PPV is defined as the fraction of positively identified cases that are true positive: number of true positives/(number of true positives + number of false positives). Sensitivity is defined as the true-positive rate: number of true

positives/(number of true positives + number of false negatives), whereas specificity is the true-negative rate: number of true negatives/(number of true negatives + number of false positives).

RESULTS

We present results of the asthma classification using three different scenarios. Each can be used to meet the requirement of a particular study. In the first scenario (Figure 2), only definite asthma cases were considered relevant for the study. The probable and doubtful asthma cases were ignored. In the second scenario (Figure 3), the definite and probable asthma cases were considered relevant for the study. The definite and probable asthma cases were combined as positive asthma cases and doubtful cases were disregarded. In the third scenario (Figure 4), the definite, probable, and doubtful asthma cases were combined as positive asthma cases.

The sensitivity, specificity, and PPV of the classifiers using over-sampling and cross-validation methods for the three scenarios are presented in the Figures 2–4.

The first experiment with 0% over-sampling was considered as the baseline in our experiments. The classifiers showed consistent behavior during the over-sampling experiments. The specificity and PPV gradually increased, and sensitivity decreased as we increased the number of negative examples in the training set. For this particular study, we selected the model using 300% over-sampling as the final classification model because of its high sensitivity and specificity. A confusion matrix of the selected classification model is presented in Table 1. The confusion matrices for all over-sampling experiments using cross validation are presented in the Appendix.

From 1601 asthma cases (i.e., definite, probable, and doubtful), only 77 (5%) were misclassified as non-asthma cases. From 3431 non-asthma cases, 1137 (33%) were misclassified as false asthma cases.

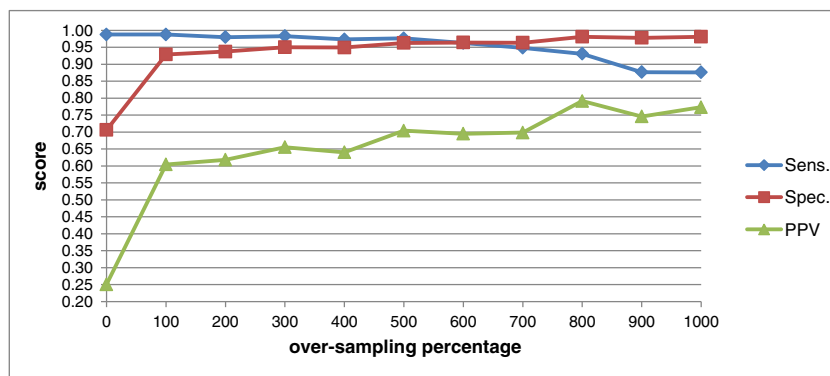


Figure 2. Performance of the classifiers using cross validation when only definite asthma cases were considered as positive asthma ignoring probable and doubtful cases

AUTOMATED CASE-DETECTION ALGORITHMS

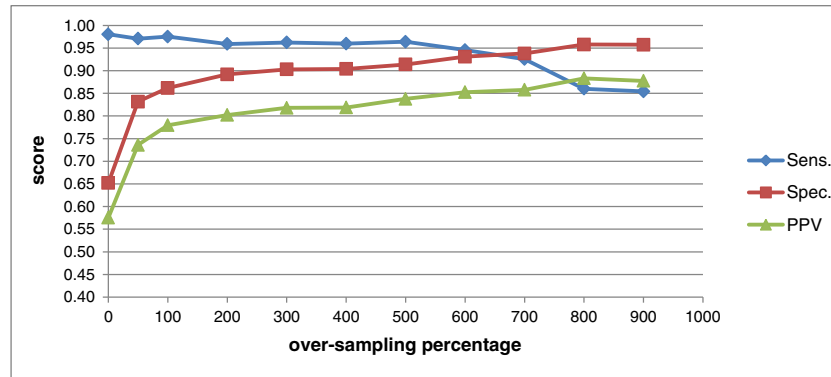


Figure 3. Performance of the classifiers using cross validation when definite and probable asthma cases were combined as positive asthma ignoring doubtful cases

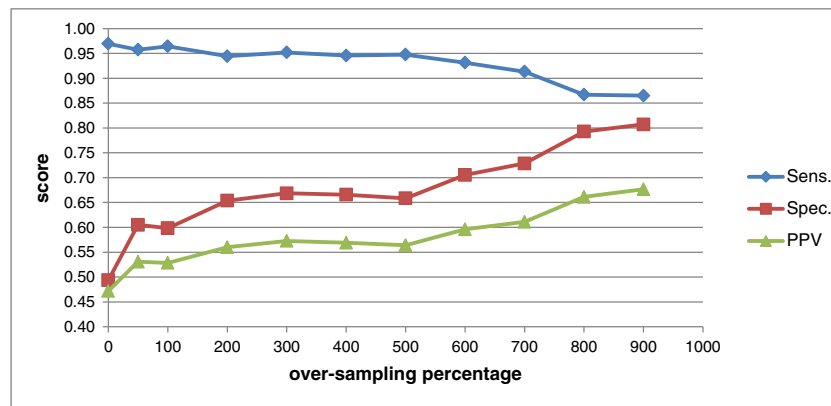


Figure 4. Performance of the classifiers using cross validation when definite, probable, and doubtful asthma cases were combined as positive asthma

Table 1. Confusion matrix of classifiers with 300 % over-sampling using cross validation

	Definite asthma	Probable asthma	Doubtful asthma	Non asthma	Patients
Definite asthma	228	47	29	4	308
Probable asthma	166	682	245	40	1133
Doubtful asthma	16	15	96	33	160
Non asthma	120	130	887	2294	3431
Total classified	530	874	1257	2371	5032

The automatic case definition for definite asthma is shown in Table 2 and for probable and doubtful asthma in Table 3.

The term “cmi” indicates an incoming communication from a specialist (i.e., a letter from a specialist). There are codes to identify specialisms (or specialists) in IPCI, and the numbers “20” and “15” are used for pediatrician and pulmonologist, respectively. Because specialists do not code events, none of the rules

Table 2. Automatic case definition for definite asthma

1. “20” and “astma” true
2. “cmi” and “astma” and not “00” and not “van” (“from”) and not “s” true
3. “cmi” and “floxotide” and not “ventolin” and not “medicatie” (“medicine”) and “20” true
4. “cmi” and “kindergeneeskunde” (“pediatrician”) and “ventolin” and not “te” (“to”) true
5. “cmi” and “astma” and “15” true
6. “cmi” and “20” and “pulmicort” true
7. “cmi” and “longziekten” (“pulmonologist”) true
8. DEFAULT false

contains an ICPC or ATC code. The drugs “floxotide,” “ventolin,” and “pulmicort” are used for obstructive airway diseases (ATC code starting with R03). Where necessary, the translation of the terms into English is included between parentheses.

The term “r03drug” indicates a respiratory drug for obstructive airway diseases. The term “r96” is the ICPC code for asthma. The terms “pulm,” “inh,” and “vag” are short for “pulmonary,” “inhaler,” and “vesiculaire ademgevoel (vesicular breath sounds),” respectively. The term “diskus” indicates a type of dry powder inhaler. Almost all rules for probable and doubtful asthma classification contain a mixture of codes and free text.

Table 3. Automatic case definition for probable and doubtful asthma

1. "r03drug" and "r96" and not "01" true
2. "r03drug" and not "hoesten" ("coughing") and not "pulm" and "flixotide" true
3. "r03drug" and not "hoesten" and not "pulm" and not "piepende" ("wheezing") and not "hoest" ("cough") true
4. "astma" and "r96" true
5. "r03drug" and not "pulm" and "inh" true
6. "ventolin" and not "pulm" and "r96" true
7. "ventolin" and "astma" and not "vag" true
8. "r03drug" and not "piepen" ("wheeze") and not "hoest" and "diskus" true
9. DEFAULT false

DISCUSSION

We created and evaluated an automated case-detection algorithm to identify children with asthma within the IPCI database. The case-detection algorithm was generated using a rule-learning algorithm that incorporated both information contained in the unstructured free-text and coded data in EMRs. We evaluated the automated algorithm in the context of three scenarios, and each scenario had different performance characteristics suitable for a different asthma study goal.

By using over-sampling techniques we could vary the performance of the resulting detection algorithm. By adding more negative examples of asthma cases, PPV and specificity increased, at the cost of decreased sensitivity (cf. Figures 2–4). Varying the amount of over-sampling allows researchers to generate a case-detection algorithm suitable for a specific study (cf. Figures 2–4); for example, when investigating incidence and prevalence, where the goal is to find the number of cases in a population in a given period, a case-detection algorithm with high sensitivity would be preferred. For our particular asthma study, we selected the algorithm with 300% over-sampling mainly because of both its high specificity and sensitivity. The selected case-detection algorithm had a PPV of 0.66, sensitivity of 0.98, and specificity of 0.95 for the scenario when only definite cases were considered relevant for the study (cf. Figure 2); PPV of 0.82, sensitivity of 0.96, and specificity of 0.90 for the scenario when definite and probable asthma cases combined were considered relevant (cf. Figure 3); and PPV of 0.57, sensitivity of 0.95, and specificity of 0.67 for the scenario when definite, probable, and doubtful asthma cases were combined and considered relevant for the study (cf. Figure 4).

Interestingly, none of the seven rules in the case-detection algorithm generated for definite asthma contains ICPC code for asthma, that is, R96 or any R03drug (cf. Table 2). The presence of the terms "flixotide," "ventolin," and "pulmicort," which are all

R03drugs, suggests that the specialists' letters do not (or not very often) contain ATC drug codes. The RIPPER algorithm was able to pick up both the terms used to indicate the specialism of pediatrics or pediatrician in the IPCI database, that is, "kindergeneeskunde" and the IPCI database code "20." Similarly, the algorithm also picked up both the terms used for the specialism of pulmonary diseases or pulmonologist, that is, "longziekten" and the IPCI database code "15." For probable and doubtful asthma cases, the algorithm picked up both the ICPC asthma code R96 and R03drug (cf. Table 3). The algorithm was also able to pick up specific drug names such as "flixotide," "ventolin," and pulmicort and abbreviations such as "inh" for "inhaler" and "vag" for "vesiculaire ademgeruis" (vesicular breath sounds) used within the IPCI database. Although some terms in the rules are hard to understand such as "van," "00," "01," "s," and "te" they would not have been picked if the RIPPER algorithm did not consider them useful or if they were not adding anything to the rules. A comparison with the broad query (Appendix) shows that the automated case definitions contain more specific keywords (and combinations) used within the database. This suggests that rules with database specific keywords are complicated to construct manually for use in the broad query.

There were some study limitations. The RIPPER algorithm used a training set of positive and negative examples of asthma cases from the IPCI database. The generated case-detection algorithm is therefore specific to the IPCI database, and it may not be applied to other databases to detect asthma cases. A new training set is required to generate an automated case-detection algorithm for a new electronic health record (EHR) database. The automated case-detection algorithm is applicable within the results of the broad query. Any asthma case initially missed by the broad query will also be missed by the automated case-detection algorithm. However, such asthma cases can potentially be identified by applying the automated case-detection algorithm onto the complete EHR database, although we do not know how well this would work.

Usually, the only way to extract or identify cases from the EHR databases is using codes such as ICPC or ICD-9 because the free-text narratives cannot be easily analyzed. Recently, Flynn *et al.*²⁷ used free-text clinical reports to develop an algorithm using manual rules to identify ischemic stroke and intracerebral hemorrhage. The approach we used in this study to generate a case-detection algorithm to identify asthma patients has a number of advantages. We used an automated approach to generate case-detection algorithms not only using the structured information

but also taking advantage of the free-text narratives present in the EHR database. We did not only present PPV but also reported sensitivity and specificity of automated case-detection algorithms. We also demonstrated how sensitivity and specificity of the algorithms can be balanced to best meet the requirements of our study. An automatic case-detection algorithm with high specificity can reduce the workload of manual annotation, by removing non-relevant records. Another advantage of automated case-detection algorithm is that they can allow for more uniform and consistent annotations as compared with several manual annotators. Although the case-detection algorithm for asthma discussed here is specific to the IPCI database, the approach used to generate the algorithm can be used in different databases.

In databases such as IPCI, manually reviewing of all results of the broad query is currently mandatory to identify asthma cases. With the use of the automated algorithm described here, it is now feasible to automatically identify definite, probable, and doubtful asthma patients with acceptable performance, using both free-text narratives and coded information when available, allowing large-scale epidemiology studies.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

KEY POINTS

- Case detection algorithms are usually created manually by using coded information such as ICPC and ICD-9 codes.
- Large amount of free-text present in most EHRs is often not used in information extraction because they cannot be easily analyzed.
- Machine learning methods can be used to generate automated case detection algorithms taking advantage of both the free-text and coded data in the EHRs.
- An automated case detection algorithm can reduce the workload of manual annotation and allow large-scale epidemiological studies.

ACKNOWLEDGEMENTS

This study was supported by the VICI project 91896632 and the Priority Medicines voor Kinderen project 113201006 of the Netherlands Organization for Health Research and Development ZonMw.

SUPPORTING INFORMATION

Supporting information may be found in the online version of this article.

Appendix:

The broad query used to identify potential asthma patients

English translation and explanation of the terms

Confusion matrices of all over-sampling experiments using cross-validation

REFERENCES

1. Bateman ED, Hurd SS, Barnes PJ, *et al.* Global strategy for asthma management and prevention: GINA executive summary. *Eur Respir J* 2008; **31**(1): 143–78.
2. Wallerstedt SM, Brunlöf G, Sundström A, Eriksson AL. Montelukast and psychiatric disorders in children. *Pharmacoepidemiol Drug Saf* 2009; **18**(9): 858–864.
3. Cates CJ, Cates MJ. Regular treatment with formoterol for chronic asthma: serious adverse events. *Cochrane Database Syst Rev* 2012; **4**: CD006923.
4. Cates CJ, Cates MJ. Regular treatment with salmeterol for chronic asthma: serious adverse events. *Cochrane Database Syst Rev* 2008; (3): CD006363. doi: 10.1002/14651858.CD006363.pub2.
5. Allen DB. Effects of inhaled steroids on growth, bone metabolism, and adrenal function. *Adv Pediatr* 2006; **53**(1): 101–110.
6. Kelly HW, Van Natta ML, Covar RA, *et al.* Effect of long-term corticosteroid use on bone mineral density in children: a prospective longitudinal assessment in the Childhood Asthma Management Program (CAMP) study. *Pediatrics* 2008; **122**(1): e53–61.
7. Incecik F, Onlen Y, Sangun O, Akoglu S. Probable montelukast-induced hepatotoxicity in a pediatric patient: case report. *Ann Saudi Med* 2007; **27**(6): 462–463.
8. Sass DA, Chopra KB, Wu T. A case of montelukast-induced hepatotoxicity. *Am J Gastroenterol* 2003; **98**(3): 704–705.
9. Linder JA, Haas JS, Iyer A, *et al.* Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting. *Pharmacoepidemiol Drug Saf* 2010; **19**(12): 1211–1215.
10. Norén GN, Hopstadius J, Bate A, Star K, Edwards IR. Temporal pattern discovery in longitudinal electronic patient records. *Data Min Knowl Discov* 2009; **20**(3): 361–387.
11. Boockvar KS, Livote EE, Goldstein N, *et al.* Electronic health records and adverse drug events after patient transfer. *Qual Saf Health Care* 2010; **19**(5): e16.
12. Persell SD, Dunne AP, Lloyd-Jones DM, Baker DW. Electronic health record-based cardiac risk assessment and identification of unmet preventive needs. *Med Care* 2009; **47**(4): 418–424.
13. Liao KP, Cai T, Gainer V, *et al.* Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res (Hoboken)* 2010; **62**(8): 1120–1127.
14. Allen-Dicker J, Klompas M. Comparison of electronic laboratory reports, administrative claims, and electronic health record data for acute viral hepatitis surveillance. *J Public Health Manag Pract* 2012; **18**(3): 209–214.
15. Schuemie MJ, Sen E, 't Jong GW, van Soest EM, Sturkenboom MC, Kors JA. Automating classification of free-text electronic health records for epidemiological studies. *Pharmacoepidemiol Drug Saf* 2012; **21**: 651–658.
16. Vlug A, van der Lei J, Mosseveld B, *et al.* Postmarketing surveillance based on electronic patient records: the IPCI project. *Methods Inf Med* 1999; **38**(12): 339–344.
17. Snowball stemmers. <http://snowball.tartarus.org/>.
18. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001; **34**(5): 301–310.
19. Setiono R, Liu H. Chi2: feature selection and discretization of numeric attributes. In *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*. IEEE Comput. Soc. Press; 1995; 388–391.
20. Kiritchenko S, Matwin S, Nock R, Famili AF. Learning and evaluation in the presence of class hierarchies: application to text categorization. Lamontagne L, Marchand M, eds *Advances in Artificial Intelligence* 2006; **4013**: 395–406.
21. Costa E, Lorena A, Carvalho A, Freitas A, Holden N. Comparing Several Approaches for Hierarchical Classification of Proteins with Decision Trees. In *Advances in Bioinformatics and Computational Biology*, Sagot M-F,

- Walter M (eds). Lecture Notes in Computer Science, vol. **4643**, Springer: Berlin, Heidelberg, 2007; pp. 126-137. doi: 10.1007/978-3-540-73731-5_12, isbn: 978-3-540-73730-8, http://dx.doi.org/10.1007/978-3-540-73731-5_12
22. Metz J, Freitas AA, Monard MC, Cherman EA. A study on the selection of local training sets for hierarchical classification tasks. In *Brazilian National Meeting on Artificial Intelligence (ENIA-2011)*. Sociedade Brasileira da Computa - SBC: Natal, RN, Brasil, 2011; 572-583.
23. Chawla NV. Data mining for imbalanced datasets: an overview. In *Data Mining and Knowledge Discovery Handbook*, Maimon O, Rokach L (eds.). Boston, MA: Springer US, 2010; 875-886.
24. Cohen WW. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, Prieditis A, Russell S (eds). Morgan Kaufmann: Tahoe City, California, USA, 1995; 115-123.
25. Geurts P, Irtum A, Wehenkel L. Supervised learning with decision tree-based methods in computational and systems biology. *Mol Biosyst* 2009; **5**(12): 1593-1605.
26. Adler W, Brenning A, Potapov S, Schmid M, Lausen B. Ensemble classification of paired data. *Comput Stat Data Anal* 2011; **55**(5): 1933-1941.
27. Flynn RWV, Macdonald TM, Schembri N, Murray GD, Doney ASF. Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes. *Pharmacoepidemiol Drug Saf* 2010; **19**(8): 843-7.

Appendix:

The broad query used to identify potential asthma patients.

"astma" OR
"asthma" OR
"wheez" OR
("piep" and not "piepen-" and not "geen" and not "niet") OR
"R96." OR
("kort" and "adem") OR
("dyspn" and not "dyspnoe-" and not "geen")

English translation and explanation of the terms

wheez	Short for wheezing
piep	Short for wheezing "cough"
geen	"no"
niet	"not"
R96	ICPC code for asthma
kort	"short"
adem	Short for dyspnea
dyspn	Short for dyspnea "shortness of breath"

Confusion matrices of all over-sampling experiments using cross-validation

0%	Definite asthma	Probable asthma	Doubtful asthma	Non asthma	Patients
Definite asthma	236	50	19	3	308
Probable asthma	326	630	155	22	1133
Doubtful asthma	39	22	75	24	160
Non asthma	705	212	820	1694	3431
Total classified	1306	914	1069	1743	5032

100%	Definite asthma	Probable asthma	Doubtful asthma	Non asthma	Patients
Definite asthma	240	46	19	3	308
Probable asthma	184	711	211	27	1133
Doubtful asthma	16	15	102	27	160
Non asthma	157	177	1044	2053	3431
Total classified	597	949	1376	2110	5032

200%	Definite asthma	Probable asthma	Doubtful asthma	Non asthma	Patients
Definite asthma	243	37	23	5	308
Probable asthma	197	646	247	43	1133
Doubtful asthma	16	10	93	41	160
Non asthma	150	127	911	2243	3431
Total classified	606	820	1274	2332	5032

300%	Definite asthma	Probable asthma	Doubtful asthma	Non asthma	Patients
Definite asthma	228	47	29	4	308
Probable asthma	166	682	245	40	1133
Doubtful asthma	16	15	96	33	160
Non asthma	120	130	887	2294	3431
Total classified	530	874	1257	2371	5032

400%	Definite asthma	Probable asthma	Doubtful asthma	Non asthma	Patients
Definite asthma	217	56	29	6	308
Probable asthma	146	696	250	41	1133
Doubtful asthma	10	11	99	40	160
Non asthma	122	125	900	2284	3431
Total classified	495	888	1278	2371	5032

500%	Definite asthma	Probable asthma	Doubtful asthma	Non asthma	Patients
Definite asthma	207	65	31	5	308
Probable asthma	136	712	248	37	1133
Doubtful asthma	9	14	95	42	160
Non asthma	87	130	955	2259	3431
Total classified	439	921	1329	2343	5032

600%	Definite asthma	Probable asthma	Doubtful asthma	Non asthma	Patients
Definite asthma	205	58	37	8	308
Probable asthma	141	654	285	53	1133
Doubtful asthma	10	10	91	49	160
Non asthma	90	93	828	2420	3431
Total classified	446	815	1241	2530	5032

700%	Definite asthma	Probable asthma	Doubtful asthma	Non asthma	Patients
Definite asthma	220	45	31	12	308
Probable asthma	143	610	310	70	1133
Doubtful asthma	10	12	81	57	160
Non asthma	95	74	762	2500	3431
Total classified	468	741	1184	2639	5032

800%	Definite asthma	Probable asthma	Doubtful asthma	Non asthma	Patients
Definite asthma	201	47	45	15	308
Probable asthma	101	580	316	136	1133
Doubtful asthma	8	6	84	62	160
Non asthma	53	70	588	2720	3431
Total classified	363	703	1033	2933	5032

900%	Definite asthma	Probable asthma	Doubtful asthma	Non asthma	Patients
Definite asthma	185	52	45	26	308
Probable asthma	116	549	340	128	1133
Doubtful asthma	3	9	86	62	160
Non asthma	63	63	536	2769	3431
Total classified	367	673	1007	2985	5032

1000%	Definite asthma	Probable asthma	Doubtful asthma	Non asthma	Patients
Definite asthma	191	47	43	27	308
Probable asthma	112	471	374	176	1133
Doubtful asthma	4	4	77	75	160
Non asthma	56	39	450	2886	3431
Total classified	363	561	944	3164	5032