Abstract

Asthma is the most common chronic childhood disease and has seen increasing prevalence worldwide. While there is existing evidence of familial and other risk factors for pediatric asthma, there is a need for further studies to explore and understand interactions among these risk factors. The goal of this study was to develop an approach for mining, visualizing, and evaluating association rules representing pairwise interactions among potential familial risk factors based on information documented as part of a patient’s family history in the electronic health record. As a case study, 10,260 structured family history entries for a cohort of 1,531 pediatric asthma patients were extracted and analyzed to generate family history associations at different levels of granularity. The preliminary results highlight the potential of this approach for validating known knowledge and suggesting opportunities for further investigation that may contribute to improving prediction of asthma risk in children.

Introduction

Asthma is the leading chronic condition in children globally and has seen a dramatic increase in prevalence over the last few decades. Several reports have described the need to better understand the effect of genetic, environmental, and lifestyle factors and their interactions for improving pediatric asthma management. With respect to familial risk factors, a number of studies have focused on exploring family history of asthma or atopic disease as a potential predictor of asthma risk in children. These studies have primarily involved use of questionnaires and examined the effect of asthma in first- and second-degree relatives (parents, siblings, and grandparents) and in multiple relatives. One study described the low positive predictive value of a family history of asthma; however, it did conclude that this could be a useful predictor for some preventive health efforts and further studies are needed to understand the role of family history in risk assessment for pediatric asthma.

The electronic health record (EHR) offers a potentially valuable source for supporting such studies since information pertaining to potential familial risk factors is increasingly collected as part of a patient’s family history in addition to patient-specific diseases and conditions longitudinally over time. This information may be documented in multiple places throughout the EHR in structured or free-text form, thus requiring the development of approaches to extract, standardize, and integrate the information for subsequent analysis. To date, there have been some efforts focused on improving documentation and use of family history in the EHR including a Stage 2 Meaningful Use measure specifying structured data entry for one or more first-degree relatives in more than 20% of patients, specifications and models for representing family history, and development of natural language processing (NLP) techniques for extracting family history from clinical notes.

Knowledge discovery and data mining approaches have the potential to transform EHR data into actionable disease knowledge. Such approaches have been used to study a variety of disease relationships (e.g., disease-disease and disease-drug) using both structured data (e.g., from problem and medication lists) and free-text within clinical notes from the EHR. Association rule mining is a commonly used data mining technique for discovering “interesting” relationships between items (e.g., family history) in large datasets (e.g., from the EHR) that could contribute to generating hypotheses for further investigation. An association rule is represented as \( X \Rightarrow Y \) (e.g., if patient has a family history of \( X \), patient also has a family history of \( Y \)) where \( X \) is referred to as the antecedent or left-hand-side (LHS) of the rule and \( Y \) is referred to as the consequent or right-hand-side (RHS) of the rule. A number of “interestingness” measures (e.g., support, confidence, lift, chi-square, and odds ratio) can be calculated to convey the strength of a given rule. Typically, minimum thresholds are specified for support (e.g., proportion of patients...
with a family history of $X$ and $Y$ relative to all patients) and confidence (e.g., proportion of patients with a family history of $X$ and $Y$ relative to patients with a family history of $X$) as a mechanism for constraining results.\textsuperscript{30}

The collection of family history data in EHR systems provides an opportunity for validating known associations and for potentially discovering new knowledge of interactions among familial risk factors and diseases. The objective of this study was to develop an approach for mining and visualizing family history associations in the EHR using open-source technologies. As a case study, the approach was evaluated for a particular condition, pediatric asthma.

**Methods**

Figure 1 provides an overview of the approach used in this study that is based on the processes for Knowledge Discovery in Databases\textsuperscript{32} and Disease Knowledge Discovery\textsuperscript{23}. The four major steps involved: (1) data selection to identify a cohort of pediatric asthma patients and extract associated family history entries from the EHR, (2) preprocessing and transformation to prepare the dataset for mining and visualization at different levels of granularity, (3) data mining to generate basic statistics and association rules, and (4) interpretation and evaluation to visualize and validate the family history associations. A combination of Ruby (2.0.0) and R (3.1.2), integrated using the RinRuby Ruby gem (2.0.3), were used for the various processing and analysis tasks.

![Figure 1. Overview of Study Approach.](image)

**Data Selection**

At the University of Vermont Children’s Hospital, the Epic EHR (Epic Systems Corporation, Verona, WI) has been in use since 2009 and includes a module consisting of both structured and free-text fields for collecting family history information.\textsuperscript{33} Two of the structured fields were the focus of this study: (1) problem – selected from a list of 222 values such as “Asthma,” “Cancer,” or “*” for Other that can be further specified in a free-text comment field (not included in this study) and (2) relation – selected from a list of 22 values including “Father,” “Maternal Grandmother,” “Other” that can be further specified in the free-text comment field (not included in this study), or “Neg Hx” for indicating a negative family history of the selected problem (Table 1).

A dataset of structured family history entries was created for a cohort of pediatric asthma patients identified using the following criteria: (1) at least one encounter in 2014, (2) age 3 to <18 years old, and (3) ICD-9-CM code 493* as an encounter diagnosis in 2014 or on the problem list. The most recent set of family history entries (consisting of the problem and relation fields) associated with an encounter in 2014 were then obtained for each patient in the cohort.

**Table 1. Examples of Preprocessed and Transformed Family History Entries.**

<table>
<thead>
<tr>
<th>Problem</th>
<th>Relation</th>
<th>Comment</th>
<th>Side of Family</th>
<th>Degree of Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>Mother</td>
<td>as a child</td>
<td>maternal</td>
<td>first</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Paternal Grandfather</td>
<td></td>
<td>paternal</td>
<td>second</td>
</tr>
<tr>
<td>*</td>
<td>Brother</td>
<td>healthy</td>
<td>-</td>
<td>first</td>
</tr>
<tr>
<td>Cancer</td>
<td>Other</td>
<td>maternal great-aunt</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cystic Fibrosis</td>
<td>Neg Hx</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Light grey shading indicates excluded field or entry for this study; “-” indicates not applicable or cannot be inferred from structured relation field.
**Preprocessing and Transformation**

The full dataset was divided into two subsets for: (1) *positive family history* – entries associated with positive family history of a problem (i.e., any relation except for “Neg Hx” and (2) *negative family history* – entries associated with negative family history of a problem (i.e., relation specified as “Neg Hx”). Within these subsets, entries with “*” selected as the problem were excluded since they do not specify a particular problem in structured form (but may include information in the free-text comment field).

For the positive family history subset, four datasets were created to enable subsequent analysis at different levels of granularity:

1. Family history of problem $P$ – only includes the problem from entries for each patient (*pos_problem* dataset)
2. Family history of problem $P$ in relation $R$ – includes problem and relation from entries for each patient (*pos_problem_relation* dataset)
3. Family history of problem $P$ in side of family $S$ – includes problem and side of family (i.e., maternal or paternal) determined based on specified relation (*pos_problem_side* dataset) (Table 1, Column 4)
4. Family history of problem $P$ in degree of relationship $D$ – includes problem and degree of relationship (i.e., first or second) determined based on specified relation (*pos_problem_degree* dataset) (Table 1, Column 5)

For the negative family history subset, one dataset was created for negative family history of problem $P$ that only included the problem from entries for each patient (*neg_problem* dataset).

The five resulting datasets were transformed into a common format in preparation for the subsequent data mining step. In this format, each row includes the patient identifier (e.g., random number) and information associated with a single entry separated by the “|” character. For example, “2[Asthma]” in the *pos_problem* dataset, “2[Asthma_Mother]” in the *pos_problem_relation* dataset, “2[Asthma_maternal]” in the *pos_problem_side* dataset, and “2[Asthma_first]” in the *pos_problem_degree* dataset.

**Data Mining**

For each dataset, basic statistics were calculated for ranking problems based on frequency and prevalence (*pos_problem*, *pos_problem_relation*, and *neg_problem* datasets only for the latter). Prevalence of a family history problem was determined relative to all family history entries associated with encounters in a particular year (2014 for this study) using a formula similar to term frequency-inverse document frequency (TF-IDF) that is often used in information retrieval to reflect importance of a term relative to a document in a collection:

$$PREV(p_d) = \frac{\sum p_d}{N_d} \times log \left( \frac{N_y}{\sum p_y} \right)$$

where $p_d$ represents patients with a family history of problem $p$ in a cohort for a particular disease $d$ (pediatric asthma), $N_d$ is the total number of patients in the disease cohort, $N_y$ is the number of patients with family history entries in year $y$ (e.g., 2014), and $p_y$ is the number of patients with a family history of problem $p$ in that year.

Association rule mining was performed using the *arules* R package (1.1-6) that interfaces with a C implementation of the Apriori algorithm. With *arules*, minimum thresholds for support and confidence as well as maximum length of rules can be specified. Numerous other interestingness measures such as lift, chi-square ($\chi^2$), and odds ratio can also be calculated for further filtering and ranking of the generated rules. Analysis of each dataset was performed using different combinations of minimum support values (0.0 to 0.1 in 0.01 increments) and minimum confidence values (0.0 to 1.0 in 0.1 increments). The maximum rule length was restricted to two in order to focus on pairwise associations. In addition to support and confidence, $\chi^2$ was also calculated and used to rank rules as this measure was found to outperform other measures in prior studies.

**Interpretation and Evaluation**

To facilitate interpretation, the *aruleVis* R package (1.0-0) was used that implements ten visualization techniques for exploring association rules generated by *arules*. Three types of visualizations were selected for viewing all rules or subsets of rules (i.e., top 50) in each dataset: (1) scatter plot that highlights the distribution of rules relative to specified measures (e.g., support, confidence, and $\chi^2$), (2) graph-based that displays vertices representing items or itemsets and edges representing relationships in rules, and (3) grouped matrix-based that uses a balloon plot for
displaying rule antecedent groups as columns and rule consequents as rows. For the latter two visualizations, support and \( \chi^2 \) were specified as the measures for depicting rule strength based on color and size of nodes respectively. For larger sets of rules (i.e., all), arulesViz was used to export rules to a GraphML format for interactive visualization using tools such as Gephi\textsuperscript{37}.

As an initial validation of the results, several literature sources were reviewed for known familial risk factors and comorbidities, including chapters on asthma within pediatric textbooks\textsuperscript{39–46} and published studies or reports focused on pediatric asthma\textsuperscript{8–16, 41}. Two pediatric clinical experts (RCW and PTR) also provided further interpretation of the family history associations.

**Results**

Using the specified criteria, 2,048 pediatric asthma patients were identified where 882 (43.1\%) were from problem list only, 68 (3.3\%) from encounter diagnosis only, and 1,098 (53.6\%) from both. Of these patients, 1,646 (80.4\%) had at least one family history entry since 2009 and 1,531 (74.8\%) had entries associated with encounters in 2014; this latter set formed the cohort for this study. Of the 10,260 most recent family history entries for this cohort (as of December 31, 2014), 7,722 (75.3\%) were associated with positive family history and 2,538 (24.7\%) with negative family history. After excluding entries with “*” indicated as the problem, 7,342 (71.6\%) positive and 2,535 (24.7\%) negative family history entries remained. Table 2 includes the distribution of entries, age associated with the most recent set of entries, and sex for the full dataset as well as the positive and negative family history subsets.

### Table 2. Distribution of Entries, Age, and Sex for Full Dataset and Positive and Negative Family History Subsets

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Entries</th>
<th># Patients</th>
<th>Age</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Entries/Patient</td>
<td># Entries/Patient</td>
<td>Range</td>
<td>Mean±SD</td>
</tr>
<tr>
<td>Full</td>
<td>10,260</td>
<td>1,531</td>
<td>1 – 34</td>
<td>6.7±5.1</td>
</tr>
<tr>
<td>Positive</td>
<td>7,342</td>
<td>1,433</td>
<td>1 – 32</td>
<td>5.1±4.1</td>
</tr>
<tr>
<td>Negative</td>
<td>2,535</td>
<td>690</td>
<td>1 – 18</td>
<td>3.7±3.0</td>
</tr>
</tbody>
</table>

* maximum age = 17.999

The frequency and prevalence of family history problems for the pediatric asthma cohort were determined for the pos\_problem, pos\_problem\_relation, and neg\_problem datasets. Table 3 includes the top 10 family history problems, family history problems with relations, and negative family history problems ranked by prevalence.

### Table 3. Ranking of Family History Problems and Relations by Prevalence ([n] indicates ranking by frequency)

<table>
<thead>
<tr>
<th>Family History of Problem ( P )</th>
<th>Family History of Problem ( P ) in Relation ( R )</th>
<th>Negative Family History of Problem ( P )</th>
</tr>
</thead>
</table>

For each of the five datasets, pairwise association rules were generated using different thresholds for support and confidence. Figure 2 depicts the change in number of rules depending on the thresholds used for the pos\_problem dataset. Rules generated using “low” thresholds (minimum support of 0.01 and confidence of 0.1) and “intermediate” thresholds (minimum support of 0.03 and confidence of 0.3) were selected for further review.

Figure 3 includes scatter plots of the 194 and 242 rules generated using the low thresholds for the pos\_problem and neg\_problem datasets respectively. The upper left quadrant highlights rules that do not occur frequently (based on support value), but have a higher \( \chi^2 \) value. The top rules based on \( \chi^2 \) for the neg\_problem dataset involved negative family history of “Severe Sprains,” “Broken Bones,” “Collagen Disease,” “Dislocations,” and “Scoliosis” (e.g., \{Severe Sprains\} \( \Rightarrow \) \{Broken Bones\} and \{Collagen Disease\} \( \Rightarrow \) \{Dislocations\} with support=0.08 and \( \chi^2 = 352 \).
Figure 2: Number of rules for combinations of minimum support and confidence thresholds for pos_problem.

<table>
<thead>
<tr>
<th>support</th>
<th>confidence</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>18069.0</td>
<td>1960</td>
<td>196</td>
<td>148</td>
<td>132</td>
<td>119</td>
<td>112</td>
<td>108</td>
<td>105</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>0.1</td>
<td>18069.0</td>
<td>1960</td>
<td>196</td>
<td>148</td>
<td>132</td>
<td>119</td>
<td>112</td>
<td>108</td>
<td>105</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>0.2</td>
<td>18069.0</td>
<td>1960</td>
<td>196</td>
<td>148</td>
<td>132</td>
<td>119</td>
<td>112</td>
<td>108</td>
<td>105</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>0.3</td>
<td>18069.0</td>
<td>1960</td>
<td>196</td>
<td>148</td>
<td>132</td>
<td>119</td>
<td>112</td>
<td>108</td>
<td>105</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>0.4</td>
<td>18069.0</td>
<td>1960</td>
<td>196</td>
<td>148</td>
<td>132</td>
<td>119</td>
<td>112</td>
<td>108</td>
<td>105</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>0.5</td>
<td>18069.0</td>
<td>1960</td>
<td>196</td>
<td>148</td>
<td>132</td>
<td>119</td>
<td>112</td>
<td>108</td>
<td>105</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>0.6</td>
<td>18069.0</td>
<td>1960</td>
<td>196</td>
<td>148</td>
<td>132</td>
<td>119</td>
<td>112</td>
<td>108</td>
<td>105</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>0.7</td>
<td>18069.0</td>
<td>1960</td>
<td>196</td>
<td>148</td>
<td>132</td>
<td>119</td>
<td>112</td>
<td>108</td>
<td>105</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>0.8</td>
<td>18069.0</td>
<td>1960</td>
<td>196</td>
<td>148</td>
<td>132</td>
<td>119</td>
<td>112</td>
<td>108</td>
<td>105</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>0.9</td>
<td>18069.0</td>
<td>1960</td>
<td>196</td>
<td>148</td>
<td>132</td>
<td>119</td>
<td>112</td>
<td>108</td>
<td>105</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>1.0</td>
<td>18069.0</td>
<td>1960</td>
<td>196</td>
<td>148</td>
<td>132</td>
<td>119</td>
<td>112</td>
<td>108</td>
<td>105</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
</tbody>
</table>

Figure 3: Scatter plots showing the distribution of rules for pos_problem (a) and neg_problem (b) using minimum support of 0.01 and confidence of 0.1. Rules are plotted based on support and $\chi^2$ values with color reflecting confidence value (darker shading represents higher confidence value).

For the pos_problem dataset, Figure 4 includes graph-based visualizations of the 39 rules and top 50 rules generated with the intermediate and low thresholds, which demonstrate how different family history problems and rules are highlighted with varying thresholds. For example, the top rule based on $\chi^2$ in Figure 4a is $\{\text{High Cholesterol}\} \Rightarrow \{\text{Hypertension}\}$ with support=0.17 and $\chi^2=171.31$ and in Figure 4b is $\{\text{Ulcers}\} \Rightarrow \{\text{Heartburn/Reflux}\}$ with support=0.01 and $\chi^2=201.64$.

Figure 5 presents grouped matrix-based visualizations for the 9 and 31 rules generated using the intermediate thresholds for the pos_problem_relation and pos_problem_side datasets respectively while Figure 6 shows all 205 rules generated for pos_problem_degree with the low thresholds in the Gephi (0.8.2) visualization tool. These three sets of rules represent more granular associations in comparison with those from pos_problem and convey the presence of family history problems in particular relatives, side of family (maternal or paternal relative), and degree of relationship (first or second degree relative). While lower ranked, there are several rules including asthma in particular relatives or side of family such as $\{\text{Asthma}_\text{Brother}\} \Rightarrow \{\text{Asthma}_\text{Mother}\}$ with support=0.03 and $\chi^2=9.28$ in Figure 5a and $\{\text{Asthma}_\text{paternal}\} \Rightarrow \{\text{Asthma}_\text{maternal}\}$ with support=0.05 and $\chi^2=1.34$ in Figure 5b. In Figure 6, the highest ranking rule is $\{\text{Arthritis}_\text{first}\} \Rightarrow \{\text{Arthritis}_\text{second}\}$ with support=0.01 and $\chi^2=99.84$ and lowest is $\{\text{Thyroid Disease}_\text{first}\} \Rightarrow \{\text{Asthma}_\text{first}\}$ with support=0.02 and $\chi^2<0.01$.

The reviewed literature sources described several known familial risk factors for pediatric asthma: family history of atopic diseases (asthma, allergic rhinitis, and atopic dermatitis [eczema]) in general as well as in specific relatives (parents, siblings, and grandparents)\(^6,8-10,38-41\). Some studies also reported on the effect of number of relatives (asthma in one or more relatives) as well as sex (male or female relatives)\(^6,10\). Asthma co-morbidities include: sinusitis, gastroesophageal reflux disease (GERD), obesity, psychological disturbances (particularly depression and...
anxiety disorders), and tobacco use. In addition, conditions reported to be possibly increased with asthma include: hypertension, diabetes, heart disease, arthritis, and cancer. Collectively, the results in Table 3 and Figure 4-6 are consistent with the existing evidence of familial risk factors and co-morbidities for asthma.

**Figure 4:** Graph-based visualization of rules for pos_problem with minimum support of 0.03 and confidence of 0.3 [39 rules] (a) and minimum support of 0.01 and confidence of 0.1 [194 rules; only top 50 shown] (b). In this representation, family history problems and rules are represented as vertices and edges indicate the relationship between problems in rules. The size of the vertex represents the $\chi^2$ value (larger circle for higher value) and color corresponds with support value (darker shade for higher value).

**Figure 5:** Grouped matrix-based visualizations of rules for pos_problem_relation (a) and pos_problem_side (b) using minimum support of 0.03 and confidence of 0.3. The x-axis displays items in the LHS of the rule and y-axis displays items in the RHS of the rule. Circle size and color reflect the strength of a rule based on $\chi^2$ and support respectively.
Figure 6: Visualization of rules in Gephi tool (using a force-directed layout algorithm, Fruchterman-Reingold) for *pos_problem_degree* using minimum support of 0.01 and confidence of 0.1 [205 rules].

**Discussion**

In this study, the feasibility of using structured family history information in the EHR was explored for identifying and assessing interactions among potential familial risk factors for pediatric asthma. The preliminary results highlight the potential of the developed approach for validating known knowledge and suggesting opportunities for further investigation. As reflected in the analyses (Figures 4-6), rules including family history of asthma, allergic rhinitis, and allergies in general, specific relatives, maternal and paternal side of family, and first- and second-degree relatives support what has been reported in the literature\(^6,8-10,38-41\). Other rules reflect reported asthma co-morbidities (e.g., Heartburn/Reflux, Depression, and Anxiety Disorders) as well as common chronic conditions in adults (e.g., Cancer, Diabetes, Heart Disease, Hypertension, and High Cholesterol)\(^{41,42}\). There is also some evidence of associations for asthma and potentially related conditions (e.g., otitis media) with Migraines\(^43,44\) and Hearing Loss\(^45,46\), which appeared in Figure 6. The findings from this study represent a first step towards further understanding the potential effect of family history of asthma, asthma co-morbidities, and other chronic conditions that may ultimately contribute to informing enhancements to tools for predicting increased risk of asthma in children (e.g., the Asthma Predictive Index\(^47\) that specifies parental asthma among the major criteria).

In performing the case study for pediatric asthma, several next steps were identified for enhancing the knowledge discovery process. Cohort identification was based on documentation of particular ICD-9-CM codes as the encounter diagnosis or on the problem list. Future studies would explore use of other data in the EHR (e.g., medications such as inhaled corticosteroids and clinical notes) and compare the accuracy of these different sources. Additional fields from the family history module in the Epic EHR at the University of Vermont Children’s Hospital as well as family history documented within clinical notes could serve to complement the structured problem and relation fields. For example, as shown in Table 1, the free-text comment field may include problems, relations, or other information that could not be documented in structured format\(^33\). Existing NLP tools for family history could be adapted to extract information from the comments and notes for subsequent integration and data mining\(^17-20\).

Several known issues are associated with association rule mining such as the generation of large numbers of rules that can present challenges for identifying meaningful associations\(^48\). A related challenge is ensuring that important rules are not missed due to high thresholds (referred to as the “rare item problem”\(^49\)). To address these challenges, a number of algorithms and techniques have been proposed such as generalized association rule mining that involves use of concept hierarchies to generate rules at different levels of granularity\(^50\). In this study, family history entries
were analyzed at four levels of granularity where the most general rules were generated from pos\_problem, most specific rules from pos\_problem\_relation, and intermediate rules from pos\_problem\_side and pos\_problem\_degree. Next steps include exploring use of existing concept hierarchies for relations (e.g., HL7 Vocabulary for RoleCode that includes almost 150 values organized in a six-level hierarchy\textsuperscript{31}) and disease groupings to also enable generalization of problems (e.g., Clinical Classifications Software\textsuperscript{52} and PheWAS groups\textsuperscript{53} for ICD-9-CM). Different thresholds for support and confidence were used to understand the effect of these constraints on number and content of rules and two sets of thresholds were selected as a demonstration (Figure 2). Further work is needed to determine the balance between number and quality of rules to facilitate interpretation and evaluation.

In interpreting the generated rules, questions arose regarding the appearance of common chronic conditions and whether there may be bias due to documentation practices. Different control populations (e.g., pediatric patients who do not have asthma or have other chronic conditions such as diabetes) could be used to compare family history associations, filter common rules, and thus highlight those that may be unique and relatively more common to the pediatric asthma population. Documentation by different providers or clinics (e.g., primary care vs. specialist) as well as variable EHR user interfaces for family history also warrants further investigation. For example, the top rules generated for neg\_problem included negative family history problems that were often associated with encounters in a pediatric orthopedics clinic, speaking perhaps to differences in documentation practices and possibly how family history questions are asked. For the initial validation, a number of literature sources were used to identify known familial risk factors and co-morbidities for asthma. Biomedical literature (e.g., in PubMed/MEDLINE) and other data sources (e.g., public health surveys) could serve as potentially valuable sources for mining and discovery of associations among risk factors for diseases that could complement or be used to validate those generated from the EHR\textsuperscript{29, 54, 55}. Next steps also include exploring additional visualization techniques and tools (e.g., R packages for heatmaps) to improve readability and facilitate interpretation as well as conducting more formal evaluations that would involve categorization of rules by clinical experts (e.g., as known/unknown or direct/indirect).

The approach used in this study was designed to be generalizable to other institutions, risk factors, and conditions. Use of open-source technologies (R and Ruby) also provides a flexible, configurable, and extensible framework. Planned extensions to the approach include incorporating standards to promote knowledge sharing and comparison across institutions. For example, the local codes assigned to problems and relations used in this study could be mapped to the Unified Medical Language System (UMLS) Metathesaurus\textsuperscript{56} and HL7 Vocabulary for RoleCode\textsuperscript{51} respectively. Based on a preliminary mapping, ~93\% of the problems could be mapped to UMLS concepts using a combination of MetaMap\textsuperscript{57} and manual searches while 73\% of the relations could be mapped to HL7 Vocabulary codes. Other next steps include incorporating risk factors documented as part of the social history in the EHR to gain insights to interactions among familial, social, and behavioral factors for diseases such as pediatric asthma\textsuperscript{58, 59}.

Conclusion

The widespread adoption of EHR systems has the potential to contribute to enhancing knowledge of interactions among risk factors for diseases. This study demonstrated the use of structured family history information from an EHR to identify pairwise associations representing interactions among potential familial risk factors for pediatric asthma. The preliminary findings support existing evidence and provide guidance for next steps in exploring the use of family history in disease risk assessment.

Acknowledgments

Research reported in this manuscript was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011364. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References