Clinical subphenotyping of asthma patients in the Severe Asthma Research Program (SARP)

Wei Wu, PhD
Lane Center for Computational Biology
School of Computer Science
Carnegie Mellon University
- Joint work with Sally Wenzel, MD at UPMC, and the SARP team
- *Journal of Allergy and Clinical Immunology, 133*(5):1280-1288
Background

- Asthma is a heterogeneous chronic airway disorder

- International and national guidelines suggest phenotyping by severity based on **lung function, asthma symptoms, and use of medication**.

- Traditional severity levels per Severe Asthma Research Program (SARP):
  - 0: normal control
  - 1: mild
  - 2: mild + ICS
  - 3: moderate
  - 4: moderate + ICS
  - 5: severe asthma
Background

- Additional factors, including inflammatory features and environmental triggers, also contribute to asthma heterogeneity

- Heterogeneity in asthma makes research and treatment difficult
  - Makes weak genetic signals even more difficult to detect
Our tasks

- To better define asthma
  - Identify new subtypes of asthma

- To better characterize patients in different subtypes of asthma, and thus make personalized medicine possible
What is the problem with the traditional definition of asthma?
Diagnosing Asthma Using Traditional Clinical Criteria

30 Subjects

0: Normal Subjects (5)
1: Mild Asthma (5)
2: Mild Asthma + Medication (5)
3: Moderate Asthma (5)
4: Moderate Asthma + Medication (5)
5: Severe Asthma (5)
Clinical profiles of asthma patients

30 patients with 112 variables

- Multiple Lung Function
- Quality of Life
- Symptom, Health Care
- Allergy
- Immune factors, Environmental factors, History, Genetics

Legend:
- Low
- Average
- High
Our Strategy

- Selected 378 patients in the SARP Program
- 112 variables were available for these patients, including inflammatory measures
- Clustering patients using unsupervised learning approach
Challenges in clustering asthma patients

- Data collection

  - Collecting data from patients can be very time consuming
    - They need to fill out questionnaires, provide blood samples, take all kinds of clinical tests ...

  - The SARP data came from 9 clinical centers, in US and UK
Data processing

- Dealing with mistakes in the database
  - Data entry errors, caused by patients, nurses, data management people, etc.
  - Duplicate patient records
  - Missing data: impute values for missing data

![Cluster (All Patients) vs Age Asthma Onset (age) chart]

Normal controls with age of asthma onset at ~ 15, 18
Finding right computational algorithms to analyze the data

- Usually clustering algorithms are designed for continuous data

- Mixed data types
  - Continuous variables, e.g., lung function variables, BMI
  - Categorical ordinal variables, e.g., symptoms
  - Categorical binary variables, e.g., whether take a certain medication or not
  - Categorical nominal variables, e.g., race, ethnicity

- Questions:
  - Which clustering algorithms to use?
  - What distance measure to use?
Importance of having a multidisciplinary team

- In order to address the challenges, computational biologists and clinicians need to work closely together
  - Modern medicine is impossible without help from computational biologists or statisticians
  - Clinicians can help with data analysis ...
    - Spotting data errors
    - Connecting your results to clinical practice
An example: spotting data errors by clinicians

- Clustering analysis revealed a small cluster of patients (5-10) with high blood neutrophil counts

<table>
<thead>
<tr>
<th>WBC_Blood</th>
<th>neutrophil pct_Blood</th>
<th>neutrophil_Blood</th>
<th>correct values</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>67.3</td>
<td>30</td>
<td>3.0</td>
</tr>
<tr>
<td>5.7</td>
<td>58.7</td>
<td>33</td>
<td>3.3</td>
</tr>
<tr>
<td>8.1</td>
<td>68.3</td>
<td>55</td>
<td>5.5</td>
</tr>
<tr>
<td>6.9</td>
<td>65</td>
<td>45</td>
<td>4.5</td>
</tr>
<tr>
<td>6.2</td>
<td>53</td>
<td>33</td>
<td>3.3</td>
</tr>
<tr>
<td>5.6</td>
<td>56.5</td>
<td>32</td>
<td>3.2</td>
</tr>
</tbody>
</table>

- Errors caused by data entry
- Can be corrected
How to evaluate clustering results?
“... In general, the selection of “good” variables is a nontrivial task and may involve quite some trial and error (in addition to subject-matter knowledge and common sense). In this respect, cluster analysis may be considered an exploratory technique.”

How to explore?

- Often time, no unique “best” or “true” clusters

- Our evaluation criterion:
  - Good clustering results should make good clinical sense
With our evaluation criterion in mind ...

- Our results make the best clinical sense when we use:
  - K-means clustering
  - Euclidean distance:
  - Set cluster number: 6 (clusters)
K-means clustering

- Given a set of data points, K-means clustering aims to minimize the within-cluster sum of squares:

\[ \arg\min_C \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2 \]

- Euclidean distance:

\[ d_E(x_i, x_j) = \sqrt{\sum_{v=1}^{p} d_v(x_{iv}, x_{jv})^2} \]

- Cluster numbers:
  - Both K = 5 and K = 6 generate results which look nice by statistical criteria
  - With K = 6, results make better clinical sense
    - Clusters are clinically recognizable
K-means Results: Patient clustering

Cluster 1

Cluster 2

Cluster 3

Cluster 4

Cluster 5

Cluster 6
Interpreting clustering results
Comparison of Patients in 6 Clusters: Lung Function

Baseline Lung Functions:

Forced Expiratory Volume in One Second: FEV1

P < 0.0001
Comparison of Patients in 6 Clusters: Improvement of lung function after treatment

Lung Function after drug treatment:

![Diagram showing comparison of patients in 6 clusters with maxFEV1pp_MPVLung data and indicating P < 0.0001]
Comparison of Patients in 6 Clusters: 
**Asthma Symptoms**

**Shortness of Breath Frequency**

- **P < 0.0001**

**Cough Frequency**

- **P < 0.0001**
Comparison of Patients in 6 Clusters: Medication

Use Oral CSs

P < 0.0001
Comparison of Patients in 6 Clusters: Health Care Utilization

Visited ER for breathing in last year

ICU admission due to asthma

P < 0.0001
Do these clusters make good clinical sense?
Patient Clusters

Patient Cluster 3: **Good lung function, bad symptoms**

- **Baseline_preDrug_FEV1pp**
  - % predicted
  - P < 0.0001

- **maxFEV1pp_MPVLung**
  - % predicted
  - P < 0.0001

- **Blood eosinophil numbers**
  - Measured Value
  - P < 0.0001
Patient Cluster 3: **Good lung function, bad symptoms**

- Clinically, a group of asthma patients tend to have depression
- This cluster of patients should probably see psychiatrists to help improve mental health
Patient clusters:

**Patient Cluster 4: Age onset/Allergy**

**Allergic Asthma**

---

**Age Asthma Onset**

- P < 0.0001

---

**Allergy symptoms in winter**

- P < 0.0001

---

**Either Mother/father with asthma**

- P < 0.0001
Patient Clusters

Patient Cluster 5: Sinus disease/eosinophilia

Had Nasal Polyp Removed

Blood eosinophil numbers

Age Asthma Onset

Nasal polyp

Asthma
Patient Clusters

Patient Cluster 6: CS/med side effect

Severe Asthma with CS side effect

Use Oral CSs

<table>
<thead>
<tr>
<th>Patient Clusters</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

P < 0.0001

Total WBC

<table>
<thead>
<tr>
<th>Patient Clusters</th>
<th>Measured Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
</tr>
</tbody>
</table>

P < 0.0001

Blood neutrophil numbers

<table>
<thead>
<tr>
<th>Patient Clusters</th>
<th>Measured Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
</tr>
</tbody>
</table>

P < 0.0001

Diagnosed with Osteoporosis

<table>
<thead>
<tr>
<th>Patient Clusters</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
</tr>
</tbody>
</table>

P < 0.0001
Can we classify the patients with respect to their cluster labels?
Supervised learning approach

Classification Strategy

Splitting patients:

- 378 patients (100%)
- 298 patients (80%)
- 80 patients (20%)

Supervised learning algorithms:
- Support Vector Machines (SVMs)
Diagnosing asthma patients with 112 variables

- Train SVM using the full set of the variables

- SVM classification accuracy rate: 95% -- highly accurate!

- But not clinically feasible!
Diagnosing asthma patients with fewer variables

- How to select fewer variables (features)?

- Feature selection using machine learning techniques:
  - Selecting informative variables using information gain
  - Selecting nonredundant variables using Markov Blanket algorithm
‘Good’ vs. ‘bad’ features

Correlated features => ‘redundant’? e.g., asthma symptom variables

Noise features => ‘less-informative’? e.g., Diagnosed with diabetes
Results from feature selection

Top informative features selected by information gain:

<table>
<thead>
<tr>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age asthma onset</td>
</tr>
<tr>
<td>Shortness of breath frequency</td>
</tr>
<tr>
<td>Activity</td>
</tr>
<tr>
<td>Symptoms</td>
</tr>
<tr>
<td>Wheeze</td>
</tr>
<tr>
<td>Chest tightness</td>
</tr>
<tr>
<td>Asthma medication use in last 3 mos</td>
</tr>
<tr>
<td>Emotion</td>
</tr>
<tr>
<td>Environment</td>
</tr>
<tr>
<td>Night-time awakenings</td>
</tr>
<tr>
<td>Asthma worsened by upper respiratory infection</td>
</tr>
<tr>
<td>Use inhaled beta-agonist</td>
</tr>
</tbody>
</table>
## Results from feature selection

- Redundant variables removed by Markov Blanket

<table>
<thead>
<tr>
<th>Nonredundant Variables</th>
<th>Redundant Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age asthma onset</td>
<td>Asthma duration, Allergy symptoms in spring, Asthma symptoms caused by animal exposure, Had allergy symptoms without cold or flu, Age, Number of skin reactions to allergens, Atopy, BMI, Blood eosinophil percentage, Mother/Father with asthma</td>
</tr>
<tr>
<td>Shortness of breath frequency</td>
<td>Wheeze, Chest tightness, Night-time awakenings, Number of children</td>
</tr>
<tr>
<td>Prebronchodilator FEV1</td>
<td>Maximum post bronchodilator FEV1, Prebronchodilator FEV1/FVC, Maximum post bronchodilator FEV/FVC, Prebronchodilator FVC, Maximal bronchodilator reversibility, Maximum post bronchodilator FVC, Diastolic BP, Blood eosinophil numbers, Number of siblings</td>
</tr>
<tr>
<td>Allergy symptoms in winter</td>
<td>Allergy symptoms in fall, Allergy symptoms in summer, BAL neutrophil percentage</td>
</tr>
</tbody>
</table>
Diagnosing asthma patients with selected variables

- 51 informative and nonredundant variables were selected by feature selection
- SVM classification accuracy rate: 88%
- However, remember, when the full set (112) of the variables was used, SVM classification accuracy rate: 95%
- Those removed features are NOT as “noninformative” and “redundant” as suggested by the statistical criteria!
How can our results so far help clinicians?
Our results can ...

- Help improve treatment of certain subtype of patients, e.g., cluster 3

- Help develop specific treatments for specific clusters of patients

  - Cluster 4: allergic asthma
  - Cluster 5: nasal polyp group
  - Cluster 6: severe asthma with CS side effect
Take home messages

- When it comes to data analysis on clinical data, details matter

  “… the application of automatic methods hoping that ‘the data will enforce its true structure’ is deceptive. Many decisions must be made for clustering.” Hennig et al., Appl. Statist (2013) 62(3): 1

- Defining phenotypes further should facilitate clinicians and researchers to:
  
  - better understand the associated pathobiology of asthma
  
  - better develop specific treatments for specific subtypes of asthma, with implications for more personalized therapy