Clustering of Chief Complaint

Jeonghoon Lee, MD\(^1\), Won Chul Cha, MD\(^{1,2,3}\)
\(^1\)Department of Digital Health, SAIHST, Sungkyunkwan University, Seoul, Korea; \(^2\)Department of Emergency Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea; \(^3\)Health Information and Strategy Center, Samsung Medical Center, Seoul, Korea

Background

Chief complaint (CC) is a starting point of diagnosis and management of patients visited Emergency Room (ER) for physicians. Since CC itself contains an important information about patients, especially in ER, it can navigate physician’s diagnosis flow. However, same CCs can be presented with different words because patients usually use variety expression to describe the main reason for visiting ER. Diversity of CC works as barrier for clinical, research area. There are some efforts to developing standard of CC and finding useful CC categorization for detecting an outbreak in syndromic surveillance, defining some patient groups. Nevertheless, availability of some sort of standardized CC is still limited for specific object.

In this circumstance, the objective of this study is to group CCs with patient’s final diagnosis through application of clustering method.

Methods

There were 734,476 ER visit records stored from 2008 to 2018 in Clinical Data Warehouse (CDW) of Samsung medical center, Korea. We removed 1,723 visit record which have no information about CC and included only patients who were older than 18 years old. Total of 654,271 cases were available for analysis. 5,789 different CCs were used during the study period, which were represented by Unified Medical Language System (UMLS) code and the number of unique UMLS codes were 1,753. To develop CC clusters, we selected diagnosis that are coded by the International Classification of Disease (ICD) codes.

To grouping of CC, first, we applied T-distributed Stochastic neighbor embedding (T-sne) technique to reduce dimentionality of CC dataset. After that CCs were clustered by means of K-means clustering method.

Results

Table 1. Top 5 most frequent cluster group

<table>
<thead>
<tr>
<th>No</th>
<th>Cluster feature</th>
<th>% of visit</th>
<th>Number of included CC</th>
<th>Top 20 Included CC (total frequency)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fever related symptoms</td>
<td>21.87%</td>
<td>44</td>
<td>fever(2,317), dyspepsia(2,319), cough(4,455), influenza(1,573), Fever with chills(2,438), COUGHED SPUTUM(1,226), Sore Throat(3,006), Dyspepsia ACUTE(2,913), Hemoptysis(2,249), General weakness(2,111), Cough with fever(1,992), Myalgia(1,476), Fever HIGH(1,129), Hyperventilation(1,180), Fatigue(1,914), Common Cold(1,523), for medication(860), PO INTOXICATION(844), DYSPIA Progressives(642), Chills(580)</td>
</tr>
<tr>
<td>2</td>
<td>Abdominal related head and facial symptoms</td>
<td>17.06%</td>
<td>60</td>
<td>Abdominal pain(5,625), Diarrhea(5,179), Fever(4,425), Nausea and vomiting NOS(4,464), Epigastric pain(3,800), Left Flank Pain(2,005), Right Flank Pain(1,715), Abdominal discomfort(1,605), ABDOMEN PAIN DIFFUS(2,823), ABDOMEN RUQ PAIN(791), ABDOMIN EPIGASTR PAIN(700), Burning epigastric pain(981), Percutaneous insertion of nephrostomy tube(516), Maintenance of drainage tube of kidney(421), Lower abdominal pain(205), Dyspepsia(315), Diarrhea and vomiting symptom(320), Abdominal Cole(522), Irrigual pain(240), Acute ear pain(169)</td>
</tr>
<tr>
<td>3</td>
<td>Trauma related head and facial symptoms</td>
<td>5.42%</td>
<td>91</td>
<td>Syncope(4,199), Neck pain(3,206), Trauma history(2,254), Traffic accidents(2,827), Facial laceration(1,952), Laceration(1,664), Scalp laceration(1,365), Facial Pain(1,203), Injury(1,677), Spitting(2,623), Laceration of forehead(598), Facial injury(573), Brain Contusion(553), Contusion of face(521), Brain Contusion(558), Contusion of scalp(502), Contusions(462), CHIN LACERATION(453), Loss of consciousness(436), Fat, NOS(193)</td>
</tr>
<tr>
<td>4</td>
<td>Neurologic symptoms</td>
<td>4.94%</td>
<td>68</td>
<td>Dizziness(2,649), Vertigo(894), Numbness(579), HAND PARESTHIESIA TINGLING(806), Gait abnormality, NOS(573), Tingling foot (hands)(222), Weakness of dital arms and legs(147), Gait abnormality(141), Dysequilibrium(118), aphasia(103), Positional vomiting(90), FINGER WAKING(554), Akasia(79), Paroxysmal(82), LOWER EXTREMTY PARESTHIESIA TINGLING(63), WEAKNESS TRANSIENT(63), Numbness in leg(62), Sensory Disorder(61), Dysasthesia(61), Paresthesia(67)</td>
</tr>
<tr>
<td>5</td>
<td>Hepatobiliary symptoms</td>
<td>4.81%</td>
<td>34</td>
<td>Achlys(1,334), Abnormal mental status(1,395), Melen(2,278), Hemorrhage(1,339), Jaundice, obstructive(2,478), Hematemesis(1,242), Abdominal distention(1,550), Abdominal distention symptom(747), Drainage Tube Care(599), Confusion(341), Impaction of colorectostomy and other biliary tube(299), Disorientation(218), Hypothermia(183), Intoxication(181), Percutaneous transhepatic insertion of biliary drain(177), Abdominal distension, gas(147), Paracenter(116), Hepatic Encephalopathy(106), Abdominal paracentesis(104), Billirubin total increased(75)</td>
</tr>
</tbody>
</table>

During study period, patients who visited ER complained of 1,753 unique CCs in 5,789 different expressions. And the number of probable UMLS codes in Samsung medical center is 36,701. The most common CC was ‘abdominal pain’ that recorded 55,625 times and the second most common CC was ‘fever’ that used 42,371 times.

We manipulated many parameters for T-sne and k-means and decided to make 30 clusters. Table 1. shows top 5 clusters that total frequency of contained CCs is high with comprised top 20 high frequency CCs.

This unsupervised CC clustering method is not a perfect way because some CCs were unacceptable included specific cluster. In spite of this result, the rest part of clusters seem like reasonably constructed with well related CCs. And the results is considered that CCs of each cluster reflect attributes of patients who suffered specific disease. For example, many of CCs that consist of 5th cluster in table1 are related with hepatobiliary system. Interestingly CCs associated with mental status are also contained in this cluster which means these CCs can be connected to ‘hepatic encephalopathy’.

From this result, we found probability of using unsupervised method to CC clustering.

Conclusions

In this study, we made 30 clusters of Chief complaints with unsupervised clustering method and most clusters consist of relational CCs. On account of diversity in expression, the effort of CC clustering could be helpful to clinical and research area. Because this way is not perfect, we are going to include more patient information and expert opinion to develop more reasonable and applicable CC clusters.