RegionAl: an Optimized Regional Classifier to Predict Mortality in Chronic Obstructive Pulmonary Disease Patients

Haohan Zhang1,2, Chunlei Tang, PhD3,4,5, Joseph M. Plasek, MS3,6, Yun Xiong, PhD1,2, Jing Ma, PhD, MD7, Li Zhou, MD, PhD3,4, David W. Bates, MD, MSc3,4,5

1Shanghai Key Laboratory of Data Science, Shanghai, CHN; 2School of Computer Science, Fudan University, Shanghai, CHN; 3Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital, Boston, MA, USA; 4Harvard Medical School, Boston, MA, USA; 5Clinical and Quality Analysis, Partners HealthCare System, Boston, MA, USA; 6Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, USA; 7Department of Population Medicine, Harvard Medical School, Boston, MA, USA.

Introduction

Studies have shown that there exists a mismatch between patients’ wishes and actual care at the end of life. In United States, up to 60% of deaths happen in an acute care hospital, with patients receiving aggressive care in their final days1 due to physicians’ tendency to over-estimate prognoses or their ability to treat the patients. For example, less than half of the 7% to 8% of all hospital admissions that need palliative care actually receive it1. Chronic Obstructive Pulmonary Disease (COPD) is a leading cause of mortality in America; affecting an estimated 14.7 million adult Americans1 diagnosed annually with COPD. Observing similar COPD conditions within a particular COPD stage is very important due to COPD’s slow progression. In this study, we define each particular COPD stage as a fixed disjoint n-day window (e.g., 30, 90, and 360 days) within the permissible tolerance range.

We previously developed a regional classifier based on a spiral timeline for visualizing literature data, which presents research topic words under different themes in a spiral map to show the chronological development of focused research topics4. When timelines are combined with a geographical map, they depict temporal patterns of events with respect to their spatial attributes4. For the task of COPD disease progression, although the regional classifier cannot be employed directly, it plays a crucial role in generating a new data product that we call a “spatial atlas of temporal variation” (abbreviated as “atlas”).

Our objective is to assess the feasibility of automatically creating visualizations of COPD disease progression to death, then to identify and track patients with palliative care needs according to a patient’s probability of death.

Materials and Methods

We extracted a cohort containing 15,500 COPD patients who had received care at Partners Healthcare network and died between 2011 and 2017 from Partners research patient data registry. Specifically, we extracted the following sections of specific types of clinical notes: PHYSICIAN INTERPRETATION from 78,489 pulmonary notes; FINDINGS and IMPRESSION from 1,893,498 “RAD/DX chest” radiology reports; and ABNORMAL ECG from 1,029,363 cardiology reports.

We propose an algorithm called Region Atlas (RegionAl), which is an optimized variant of our previously-developed regional classifier based on a spiral timeline. We focus on two aspects of optimization. First, we design an automatic procedure to identify representative sentences by extracting the initial sentences containing a fixed minimum of 30 percent of words that map to a specific topic identified by Latent Dirichlet Allocation (LDA) – the dynamic LDA model, an existing information extraction tool that was previously used in development of our regional classifier. Second, we use a tolerance range to achieve window (i.e., time interval) enlargement. Tolerance range is defined as the distance of observing similar COPD conditions between the upper and lower specification limit. Given a fixed disjoint n-day window as interval, the tolerance range is

$$\Delta_{win} = \sigma \left( \frac{e^z - e^{-z}}{e^z + e^{-z}} \right) \cdot interval$$

(1)
The function is $\sigma \cdot \tanh(z) \cdot \text{interval}$ where $\sigma = 1, 2, 3, \ldots$, $z = \text{average(size)} / \text{size}$. Note that $\sigma = 0$ means there is no tolerance range. The reasons we use the tanh function is that its outputs range from -1 to 1. The $\tanh(z) \cdot \text{interval}$ maps the relative size of text (i.e., the reciprocal of $z$) to the $[0 \cdot \text{interval}, 1 \cdot \text{interval}]$ output range because the scale of the horizontal axis is greater than or equals to zero. The $\sigma \cdot (\tanh(z) \cdot \text{interval})$ could help us modify the time interval based on different values of $\sigma$. Thus, $z$ is proportional to $\Delta \text{win}$ and inversely proportional to the relative size of texts of a specific stage: when the relative size $\rightarrow 0$, then $z \rightarrow +\infty, \Delta \text{win} \rightarrow \sigma \cdot \text{interval}$; when the relative size $\rightarrow +\infty$, then $z \rightarrow 0, \Delta \text{win} \rightarrow 0$. We know that since $\text{interval}$ is fixed with a tolerance range, it must have a blurred border in the range of $\left[\text{lower bound} = \frac{1}{2} \Delta \text{win}, \text{upper bound} = \frac{1}{2} \Delta \text{win}\right]$.

To evaluate our algorithm, we use linear regression and support vector machines (SVM) to calculate precision. On the data from both single type (e.g., only cardiology reports) and merged (i.e., when all 3 note type are included) notes, we validate tolerance test under positive predictive value (PPV) (i.e., precision), then to calculate their improvement rates (IR) by the formula as follows:

$$IR_\sigma = \frac{PPV_\sigma}{PPV_{\sigma=0}}$$

$$\text{Table 1. The average improvement rate for each value of } \sigma \text{ on the data from both single and merged note types.}$$

<table>
<thead>
<tr>
<th>Data</th>
<th>$PPV_{\sigma=0}$</th>
<th>$IR_{\sigma=1}$</th>
<th>$IR_{\sigma=2}$</th>
<th>$IR_{\sigma=3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulmonary Notes</td>
<td>5.9%</td>
<td>45.3%</td>
<td>121.8%</td>
<td>190.0%</td>
</tr>
<tr>
<td>Radiology Reports</td>
<td>9.8%</td>
<td>1.2%</td>
<td>107.7%</td>
<td>174.0%</td>
</tr>
<tr>
<td>Cardiology Reports</td>
<td>7.9%</td>
<td>0%</td>
<td>29.5%</td>
<td>113.7%</td>
</tr>
<tr>
<td>Merged Notes</td>
<td>9.3%</td>
<td>0%</td>
<td>90.5%</td>
<td>64.7%</td>
</tr>
</tbody>
</table>

Discussion and Conclusion

RegionAl is able to 1) efficiently give a probability of death for the individual patient by comparing a particular clinical document with the atlas under the permissible tolerance range; and 2) to visualize COPD conditions. This method is an improvement upon our previously developed regional classifier. The experimental results demonstrate that RegionAl produces interpretable, accurate, and reliable results in estimating COPD mortality. The atlas could also address a proxy problem – to predict the probability of death for a given patient within a permissible tolerance range, and to help make recommendations for palliative care referral. Future work will evaluate our algorithm using diverse data sets with different mortality rates and from different data resources.

References