An innovative solution for breast cancer textual big data analysis

N. Thiébaut¹, K. Neuberger¹, A. Simoulin¹, I. Ibnouaheim¹, N. Rémy¹, M. Lodig¹ and C. Mathelin¹

1. Quantmetry, 128 rue du Faubourg Saint-Honoré, Paris 75008, France
2. Hôpital de Hautepierre, 1 Avenue Molière, Strasbourg 67200, France

Work in progress, ClinicalTrials.gov Identifier: NCT02310093

Introduction

Hospitals continuously gather huge amounts of textual data in electronic health records. Storing health data in text format is convenient, but without processing, search and analysis operations on such data become tedious.

Here we present an innovative solution for the extraction of structured information out of a corpus of multidisciplinary meeting notes and hospital letters. It relies on standard text mining methods, on the NegEx [1] algorithm for negation detection and a synonym detection method based on the word2vec [2] algorithm.

Data transformation pipeline

Starting from many Microsoft Word® documents, we have used the python-docx package to obtain JSON files reminiscent of the raw documents structure. Then the JSON files collection was used as input of the pipeline depicted below.

At the end of the pipeline, automatic hypothesis testing was performed to highlight statistically significant differences between sub-populations.

Synonym detection with the word2vec algorithm

To retrieve named entities from the documents, one must be able to handle synonyms, typographical errors and acronyms. To do so, we have adopted an iterative method in which the medical staff provided an initial list of equivalent formulations for all the concepts under investigation. We could further complete this initial list with similar words found using the word2vec algorithm.

Among similar words found using the word2vec word embedding, we find synonyms as expected, but also hyponyms, hyperonyms, antonyms, and semantically related words that appear in similar contexts. These similar words were presented to the medical staff and validated before incorporation in the synonyms dictionary.

Extraction evaluation

In order to estimate the precision of our structuring methodology, we have compared the extracted values with two different log-books maintained by the hospital. Comparison with the first logbook gives encouraging preliminary results with 95% agreement.

Result example

We have used our structuring program to compare breast cancer characteristics for 7,071 patients older than 50 years old with invasive tumors and no diabetes, and 473 similar patients with type 2 diabetes. The results are shown in the table below.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type of Cancer</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average age</td>
<td>75.3</td>
<td>0.001</td>
</tr>
<tr>
<td>Baseline tumor size</td>
<td>2.3</td>
<td>0.01</td>
</tr>
<tr>
<td>Hormone receptor status</td>
<td>0.4</td>
<td>0.003</td>
</tr>
<tr>
<td>HER2 status</td>
<td>0.2</td>
<td>0.001</td>
</tr>
<tr>
<td>Grade</td>
<td>0.3</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Conclusion

Our method allows for a versatile text structuring, and can be adapted to any language. We have built lexicons with the help of the medical staff and a word2vec based synonym detection algorithm. This method can be adapted to any corpus, and is complementary to the use of clinical healthcare terminologies such as SNOMED or UMLS.

References


Reproduction policy

Preliminary results. Any reproduction, even partial, is strictly prohibited.