

LDA2Net: Digging under the surface of COVID-19 topics in literature

Topic 6 companion sheet

G. Minello

C.R.M.A. Santagiustina

M. Warglien

This file contains the following supplementary information for Topic 6 of the manuscript “*LDA2Net*: Digging under the surface of COVID-19 topics in scientific literature”:

- Human label and automatic n-gram label proposals (Table 1)
- Summary measures (Table 2)
- Network of top 25 bigrams (Figure 1)
- Wordclouds of top 25 words by node relevance measure (Figure 2)
- Wordclouds of top 25 bigrams by edge relevance measure (Figure 3)
- Filtered (0.99 percentile) topic network (Figure 4)

Table 1: Human and automatic label proposals. Automatic label candidate for largest word community of the topic. In parenthesis: absolute frequency of the walk out of a sample of size 1000.

Human label	2-gram label	3-gram label	4-gram label
retrospective clinical study	hospitalized->confirmed (5.4%)	hospitalized->confirmed->without (2.5%)	hospitalized->confirmed->without->diagnosed (0.8%)

Here follows the set of topic-specific measures that have been used to classify the topic and to analyse its structural properties (see manuscript for details):

Table 2: Summary measures

	JSD	Mean propensity	Variance propensity	Modularity	Barrat Clustering Coeff.
value	0.894972	0.009792	0.000263	0.000000	0.658041
rank	117	110	47	6	120

Based on the aforementioned measures, Topic 6 has been classified as a CROSS-CUTTING topic.

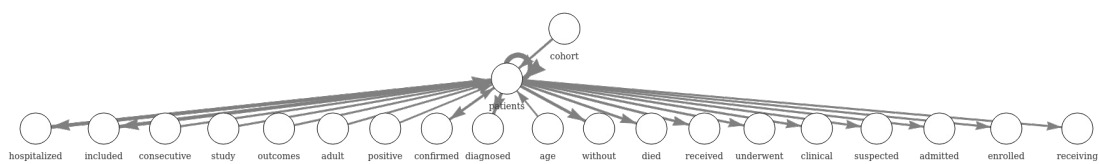


Figure 1: Network of top 25 bigrams (i.e., edges) by weight.

A word cloud of terms related to clinical research, with 'patients' as the largest word. Other prominent words include 'diagnosis', 'cohort', 'study', 'prospectively', 'retrospective', 'hospitalized', 'clinical', 'died', 'consecutive', 'total', 'confirmed', 'diagnosed', 'included', 'hospital', 'april', 'without', 'outcomes', 'patient', 'enrolled', 'age', 'median', 'suspected', 'march', 'received'.

A word cloud of medical terms centered around the word 'patients'. The words are arranged in a circular pattern, with 'patients' being the largest and most central. Other prominent words include 'diagnosed', 'hospitalized', 'admitted', 'undergoing', 'confirmed', 'received', 'enrolled', 'recovered', 'required', 'cohort', 'discharged', 'hospital', 'positive', 'clinical', 'died', 'without', 'analyzed', 'study', 'suspected', 'presented', 'underwent', 'age', 'receiving', and 'required'. The words are in various shades of gray and sizes, creating a dense, circular composition.

admitted
received
hospital
died
outcomes
clinical enrolled
study
patients
hospitalized total
diagnosed confirmed
without status
patient march
required analyzed
suspected
underwent
male
positive
age
included

A word cloud of terms related to COVID-19 research. The most prominent word is 'patients'. Other significant words include 'diagnosis', 'outcomes', 'diagnosed', 'admitted', 'study', 'clinical', 'total', 'received', 'age', 'hospitalized', 'without', 'confirmed', 'hospital', 'suspected', 'april', 'patient', 'receiving', 'enrolled', 'underwent', 'positive', 'died', 'march', and 'patient'. The words are arranged in a circular pattern, with 'patients' at the center.

PageRank

3



Figure 3: Top 25 bigrams (i.e., edges) by measure.

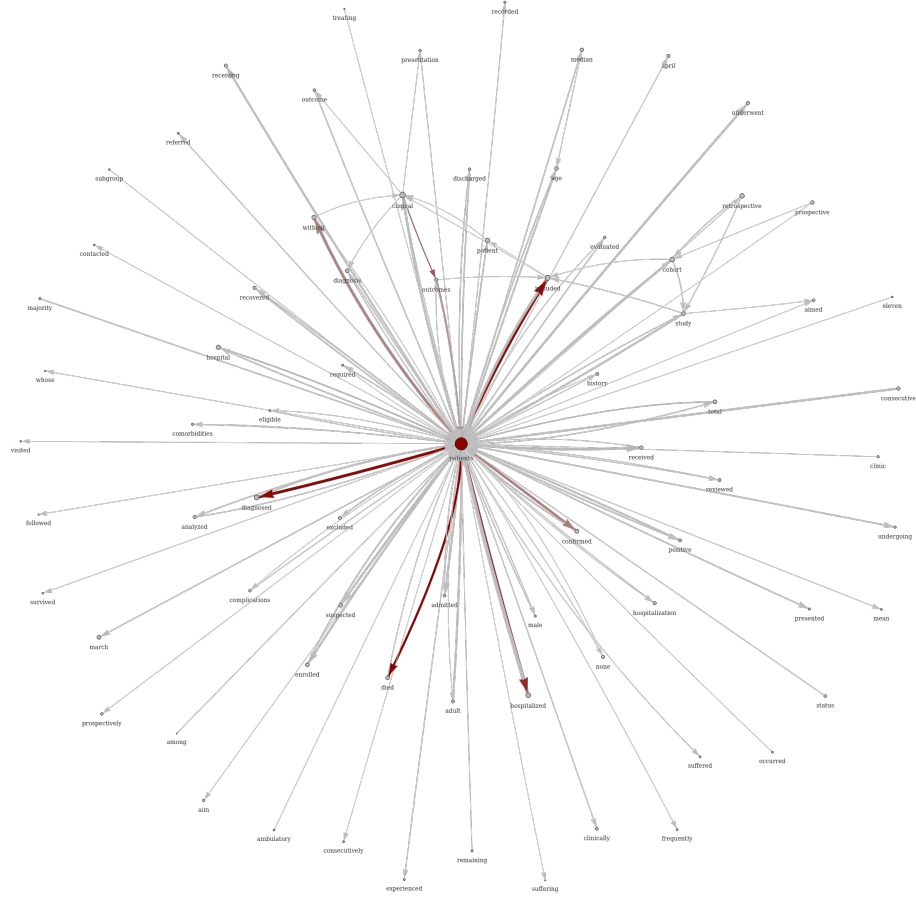


Figure 4: Filtered topic network (by weight). Layout based on Fruchterman-Reingold algorithm. Node size is proportional to topic-specific word probability provided by LDA. Edge width is proportional to topic-specific bigram weight provided by LDA2Net method. Node and edge color represent their betweenness centrality. Isolated nodes have been removed after filtration.