Understanding COVID-19 pandemic trajectories: why changes in online behavior matter for now-casting

Sara Mesquita[1], Lília Perfeito[1], João Loureiro[1], Cláudio Haupt-Vieira[2], and Joana Gonçalves-Sá[1,2,3]

[1]LIP and Physics Department, Instituto Superior Técnico, Lisboa, Portugal, [2]Nova School of Business and Economics, Carcavelos, Portugal, [3]Instituto Gulbenkian de Ciência, Oeiras, Portugal

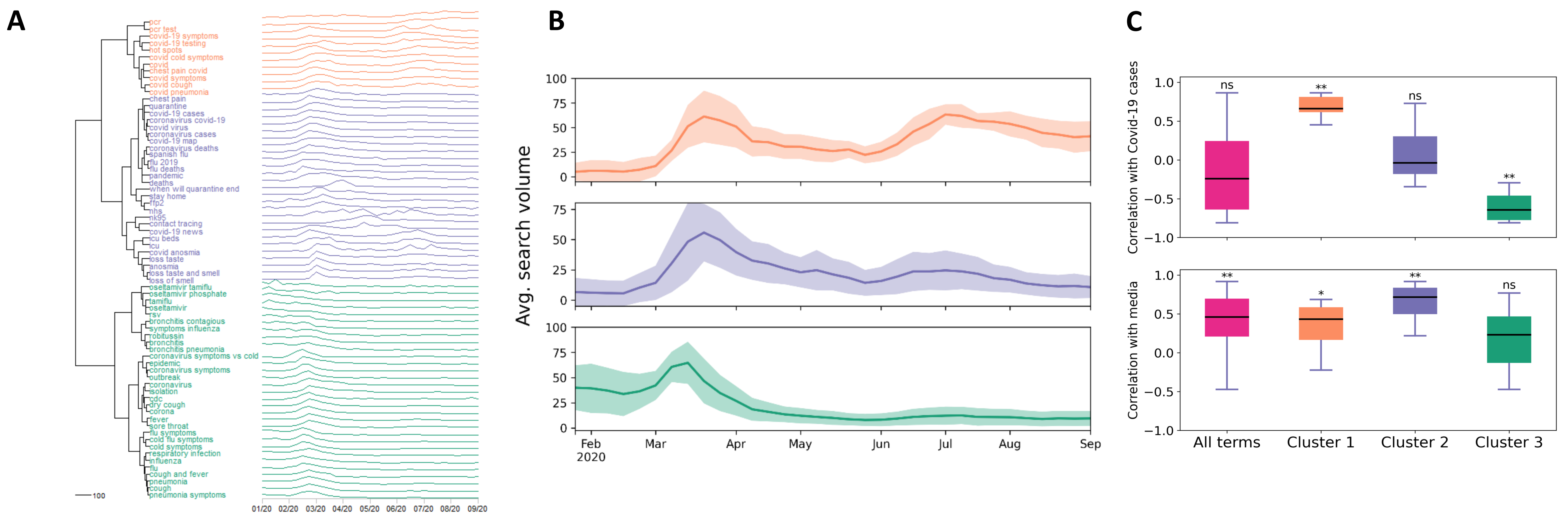
INTRODUCTION

Online behavior has been used as a tool for close to real-time study of different health-related behaviors, including identifying disease outbreaks. However, its validity in predicting infections, has been questioned by many, particularly during extraordinary events, such as the COVID-19 pandemic. Here we hypothesized that, given the increase in search volume and the co-existence of both media hype and a large number of infections right from the start of the outbreak, the earlier period of the COVID-19 pandemic could be used to disentangle between searches driven by media attention from searches driven by actual disease. The rationale is that by having periods in which media attention and cases are decoupled, it should be possible to identify which search-terms are more sensitive to media hype. This assumes that models can be improved not necessarily by blindly increasing the size of the training dataset, but by integrating prior information. Our work indicates that it is possible to learn from pandemics to improve the now-casting of future infection waves and that including more data is not necessarily better.

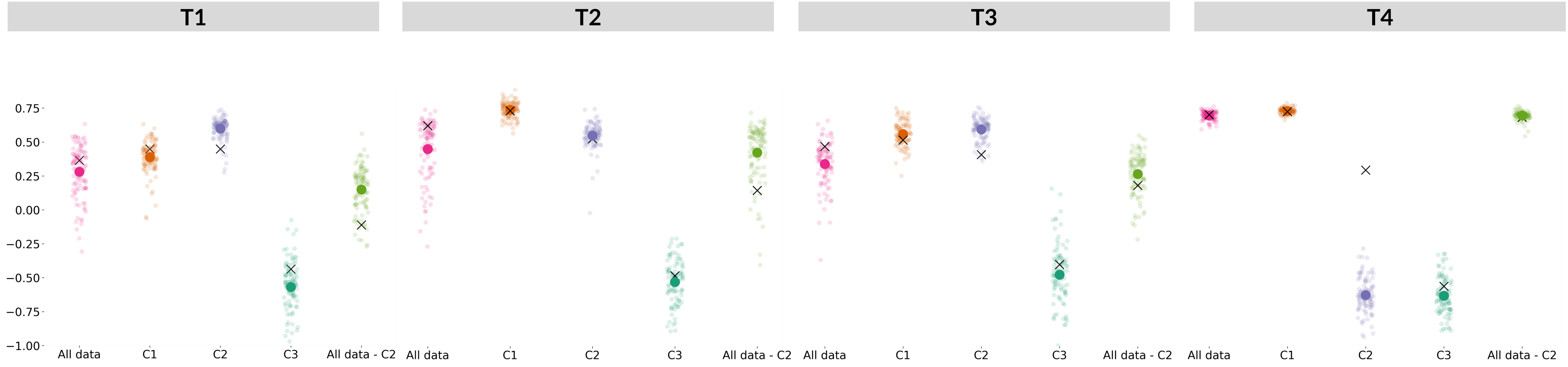
METHODS

Datasets: We collected a) Google search volume, using the GT API, selected to cover various aspects of pandemic interest, such as symptoms, antivirals, personal care measures, institutions and pandemic circumstantial terms. 69 search terms for Covid-19 in USA from Jan 2020 to October 2021. b) frequency of news media mentioning "COVID-19" or "Coronavirus" terms using Mediacloud (mediacloud.org), and c) number of actual infections. All data series were binned in a weekly resolution. **Methods**: *Clusters* - hierarchical clustering using Euclidean distance and Ward's linkage criterion. *Correlations:* Pearson correlation. *Models:* Linear Regression and Random Forest models were used to predict Covid-19 cases.

RESULTS



Different patterns of searches during COVID-19 pandemic - We found that search-terms cluster into three groups and that apparently similar terms display very different temporal dynamics: **A -** Hierarchical clustering of Google Trends search terms for Covid-19 in USA; **B -** Centroid and standard deviation over time for each cluster. **C -** Pearson correlation between the cluster centroid and either the Covid-19 cases (top) or the media mentions (bottom).* denotes 0.01 < *p-value* < 0.05, ** denotes *p-value* < 0.001, and ns a non-significant *p-value.*



Linear Regression (X) and Random Forest (●) results considering different periods - We trained different models using all the collected search terms or only the terms from C1, C2 and C3, and All data except C2, separately. We then tested their performance in nowcasting subsequent pandemic waves and found that the best accuracy is achieved when using only C1. **T1** - Model results considering the training period from January 2020 until September 2020, and testing period from September 2020 until October 2021; **T2 -** Model results considering the training period from January 2020 until September 2020, and testing period from September 2020 until March 2021; **T3 -** Model results considering the training period from January 2020 until September 2021, and testing period from September 2020 until June 2021; T4 - Model results considering the training period from January 2020 until March 2021, and testing period from March 2021 until October 2021.

CONCLUSIONS

By testing different periods, we show that C1 offers consistently better predictions than considering all the terms. In fact, by comparing the centroids of the different clusters we find that C1 correlates strongly with actual infection, and we can hypothesize that it is more disease driven, C2 search terms correlate more with media attention, and we hypothesize that it is more media driven, and C3 does not strongly correlate with either. This finding is also contrary to the common claim that more data is always better, as using a lower volume of selected data improved model accuracy, especially in the long run. Therefore, we offer three main contributions and show that: (1) Search data can be used for disease monitoring, (2) We can learn from pandemics to improve seasonal or secondary now-casting, (3) More data is not always better and including insights from behavior can improve statistical models.

REFERENCES

[1] Ginsberg, J., et. al., (2009). Detecting influenza epidemics using search engine query data. Nature, 457, 1012–4. 1

[2] Lazer, D., et. al., (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science, 343, 1203–1205. 2

[3] Towers, S.et al. (2015). Mass media and the contagion of fear: the case of ebola in america.PloS one10, e0129179

[4] Tizzoni, M., et. al., (2020). The impact of news exposure on collective attention in the united states during the 2016 zika epidemic. Plos computational biology, e1007633

[5] Kogan, Nicole E., et al.(2021) "An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time." Science Advances 7.10: eabd6989.

smesqita@lip.pt