Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset

Edwin Zhang,[1] **Nikhil Gupta,**[1] **Rodrigo Nogueira,**[1] **Kyunghyun Cho,**[2,3,4] and **Jimmy Lin**[1]

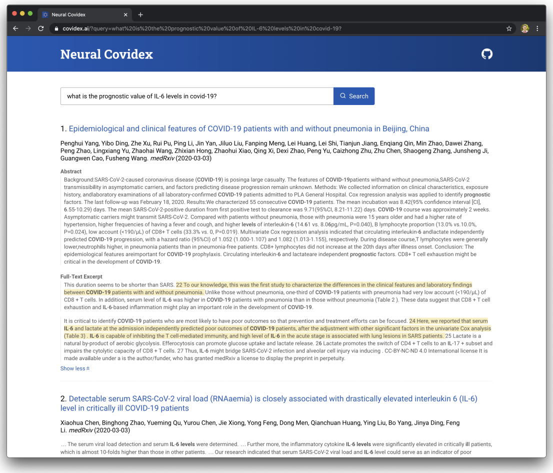
[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] Courant Institute of Mathematical Sciences, New York University
[3] Center for Data Science, New York University [4] CIFAR Associate Fellow

This extended abstract represents an abridged version of Zhang et al. (2020a), posted on arXiv April 10, 2020 and concurrently submitted to this workshop. We have intentionally decided for this short piece to reflect the state of our work at that time. The latest updates on our project can be found in Zhang et al. (2020b).

The Neural Covidex is a search engine that exploits the latest neural ranking architectures to provide information access to the COVID-19 Open Research Dataset (CORD-19) curated by the Allen Institute for AI (Wang et al., 2020). It exists as part of a suite of tools we have developed to help domain experts tackle the ongoing global pandemic. We hope that improved information access capabilities to the scientific literature can inform evidence-based decision making and insight generation.

The first version of CORD-19 was released on March 13, 2020. Within a couple of weeks, our team was able to build, deploy, and share with the research community a number of open-source components that support information access to this corpus. These include: Extensions to our Anserini IR toolkit (Yang et al., 2018) and its Pyserini Python interface (Akkalyoncu Yilmaz et al., 2020) to support basic keyword search capabilities on the corpus; PyGaggle, a new library for neural text ranking that includes supervised ranking models based on T5 as well as unsupervised sentence highlighting models with BioBERT (Lee et al., 2020).

We have assembled these components into the Neural Covidex, available online at `covidex.ai`; see screenshot in Figure 1. This user interface was developed from scratch and is itself open source. Zhang et al. (2020a) described our initial efforts and shared lessons we learned along the way.



covidex.ai/?query=what%20is%20the%20prognostic%20value%20of%20IL-6%20levels%20in%20covid-19?

Neural Covidex

what is the prognostic value of IL-6 levels in covid-19?

1. Epidemiological and clinical features of COVID-19 patients with and without pneumonia in Beijing, China

Penghui Yang, Yibo Ding, Zhe Xu, Rui Pu, Ping Li, Jin Yan, Jiluo Liu, Fanping Meng, Lei Huang, Lei Shi, Tianjun Jiang, Enqiang Qin, Min Zhao, Dawei Zhang, Peng Zhao, Lingxiang Yu, Zhaohai Wang, Zhixian Hong, Zhaohui Xiao, Qing Xi, Dexi Zhao, Peng Yu, Caizhong Zhu, Zhu Chen, Shaogeng Zhang, Junsheng Ji, Guangwen Cao, Fusheng Wang. *medRxiv* (2020-03-03)

Abstract

Background:SARS-CoV-2-caused coronavirus disease (**COVID-19**) is posinga large casualty. The features of **COVID-19**patients withand without pneumonia,SARS-CoV-2 transmissibility in asymptomatic carriers, and factors predicting disease progression remain unknown. Methods: We collected information on clinical characteristics, exposure history, andlaboratory examinations of all laboratory-confirmed **COVID-19** patients admitted to PLA General Hospital. Cox regression analysis was applied to identify **prognostic** factors. The last follow-up was February 18, 2020. Results:We characterized 55 consecutive **COVID-19** patients. The mean incubation was 8.42(95% confidence interval [CI], 6.55-10.29) days. The mean SARS-CoV-2-positive duration from first positive test to clearance was 9.71(95%CI, 8.21-11.22) days. **COVID-19** course was approximately 2 weeks. Asymptomatic carriers might transmit SARS-CoV-2. Compared with patients without pneumonia, those with pneumonia were 15 years older and had a higher rate of hypertension, higher frequencies of having a fever and cough, and higher **levels** of interleukin-**6** (14.61 vs. 8.06pg/mL, P=0.040), B lymphocyte proportion (13.0% vs.10.0%, P=0.024), low account (<190/μL) of CD8+ T cells (33.3% vs. 0, P=0.019). Multivariate Cox regression analysis indicated that circulating interleukin-**6** andlactate independently predicted **COVID-19** progression, with a hazard ratio (95%CI) of 1.052 (1.000-1.107) and 1.082 (1.013-1.155), respectively. During disease course,T lymphocytes were generally lower,neutrophils higher, in pneumonia patients than in pneumonia-free patients. CD8+ lymphocytes did not increase at the 20th days after illness onset. Conclusion: The epidemiological features areimportant for **COVID-19** prophylaxis. Circulating interleukin-**6** and lactateare independent **prognostic** factors. CD8+ T cell exhaustion might be critical in the development of **COVID-19**.

Full-Text Excerpt

This duration seems to be shorter than SARS. 22 To our knowledge, this was the first study to characterize the differences in the clinical features and laboratory findings between **COVID-19** patients with and without pneumonia. Unlike those without pneumonia, one-third of **COVID-19** patients with pneumonia had very low account (<190/μL) of CD8 + T cells. In addition, serum level of **IL-6** was higher in **COVID-19** patients with pneumonia than in those without pneumonia (Table 2). These data suggest that CD8 + T cell exhaustion and **IL-6**-based inflammation might play an important role in the development of **COVID-19**.

It is critical to identify **COVID-19** patients who are most likely to have poor outcomes so that prevention and treatment efforts can be focused. 24 Here, we reported that serum **IL-6** and lactate at the admission independently predicted poor outcomes of **COVID-19** patients, after the adjustment with other significant factors in the univariate Cox analysis (Table 3) . **IL-6** is capable of inhibiting the T cell-mediated immunity, and high level of **IL-6** in the acute stage is associated with lung lesions in SARS patients. 25 Lactate is a natural by-product of aerobic glycolysis. Efferocytosis can promote glucose uptake and lactate release. 26 Lactate promotes the switch of CD4 + T cells to an **IL**-17 + subset and impairs the cytolytic capacity of CD8 + T cells. 27 Thus, **IL-6** might bridge SARS-CoV-2 infection and alveolar cell injury via inducing . CC-BY-NC-ND 4.0 International license It is made available under a is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

2. Detectable serum SARS-CoV-2 viral load (RNAaemia) is closely associated with drastically elevated interleukin 6 (IL-6) level in critically ill COVID-19 patients

Xiaohua Chen, Binghong Zhao, Yueming Qu, Yurou Chen, Jie Xiong, Yong Feng, Dong Men, Qianchuan Huang, Ying Liu, Bo Yang, Jinya Ding, Feng Li. *medRxiv* (2020-03-03)

… The serum viral load detection and serum **IL-6 levels** were determined. … Further more, the inflammatory cytokine **IL-6 levels** were significantly elevated in critically ill patients, which is almost 10-folds higher than those in other patients. … Our research indicated that serum SARS-CoV-2 viral load and **IL-6** level could serve as an indicator of poor

Figure 1: Screenshot of the Neural Covidex.

Although the application of BERT to text ranking is well known (Nogueira and Cho, 2019), we decided to deploy our latest research based on sequence-to-sequence models (Nogueira et al., 2020), specifically T5 (Raffel et al., 2019). This relevance classifier, which reranks BM25 results from Anserini, is fed a query q and each candidate document d in turn. The model is fine-tuned to produce either "true" or "false" depending on whether the document is relevant or not to the query. At inference time, we softmax the logits of the "true" and "false" tokens, and the resulting probability of the "true" token is used as the relevance score of d. Candidate documents are then reranked using their relevance scores. As there is no COVID-19 training data, we fine-tuned our model on the MS MARCO passage dataset (Nguyen et al., 2016), and thus our reranker operates in a zero-shot setting.

As our work pre-dated any systematic evaluation efforts by the community, at the time of submission we were unable to provide any experimental results. Since then, however, we have participated in the TREC-COVID challenge (Voorhees et al., 2020); partial results to date are reported in Zhang et al. (2020b). Nevertheless, the speed at which we were able to build and deploy the Neural Covidex is a testament to the power of open-source software and modern open-science norms.

Acknowledgments

This research was supported in part by the Canada First Research Excellence Fund, the Natural Sciences and Engineering Research Council (NSERC) of Canada, NVIDIA, eBay, Microsoft, and CIFAR. We'd like to thank Kyle Lo for helpful discussions and Colin Raffel for his assistance with T5.

References

Zeynep Akkalyoncu Yilmaz, Charles L. A. Clarke, and Jimmy Lin. 2020. A lightweight environment for learning experimental IR research practices. In *Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, pages 2113–2116.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv:1611.09268*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv:1901.04085*.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv:2003.06713*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683*.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1).

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 Open Research Dataset. *arXiv:2004.10706*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: reproducible ranking baselines using Lucene. *Journal of Data and Information Quality*, 10(4):Article 16.

Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and Jimmy Lin. 2020a. Rapidly deploying a neural search engine for the COVID-19 Open Research Dataset: Preliminary thoughts and lessons learned. *arXiv:2004.05125*.

Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang, and Jimmy Lin. 2020b. Covidex: Neural ranking models and keyword search infrastructure for the COVID-19 Open Research Dataset. *arXiv:2007.07846*.