THE HONG KONG
POLYTECHNIC UNIVERSITY

Extreme Multi-label Classification on COVID-19 Literature

Rong Xiang, Jinghang Gu

gujinghangnlp@gmail.com

10/30/2020

Linguistic Data Consortium Workshop

Outline

- Introduction
- Related Work
- Dataset
- Methodology
- Experiments & Conclusion

Multi-Label Classification for COVID-19 Semantic Indexing

Review > Eur Rev Med Pharmacol Sci. 2020 Apr;24(8):4539-4547.
doi: 10.26355/eurrev_202004_21038.

Efficacy of chloroquine and hydroxychloroquine in the treatment of COVID-19

S A Meo [1], D C Klonoff, J Akram

Affiliations + expand

PMID: 32373993 DOI: 10.26355/eurrev_202004_21038
Free article

Abstract

Objective: The novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), also called COVID-19, has caused a pandemic which has swiftly involved the entire world and raised great public health concerns. The scientific community is actively exploring treatments that would potentially be effective in combating COVID-19. Hydroxychloroquine has been demonstrated to limit the replication of SARS-CoV-2 virus in vitro. In malarial pandemic countries, chloroquine is widely used to treat malaria. In malarial non-pandemic nations, chloroquine is not widely used. Chloroquine and hydroxychloroquine share similar chemical structures and mechanisms of action. The aim of this study was to indirectly investigate the efficacy of chloroquine and hydroxychloroquine for the treatment of COVID-19 by determining the prevalence of COVID-19 in malaria pandemic and non-pandemic nations. We sought evidence to support or refute the hypothesis that these drugs could show efficacy in the treatment of COVID-19.

Materials and methods: We reviewed in vitro studies, in vivo studies, original studies, clinical trials, and consensus reports, that were conducted to evaluate the antiviral activities of chloroquine and hydroxychloroquine. The studies on "COVID-19 and its allied treatment were found from World Health

Categorization

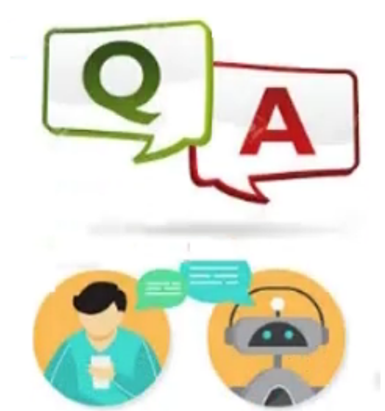
MeSH terms

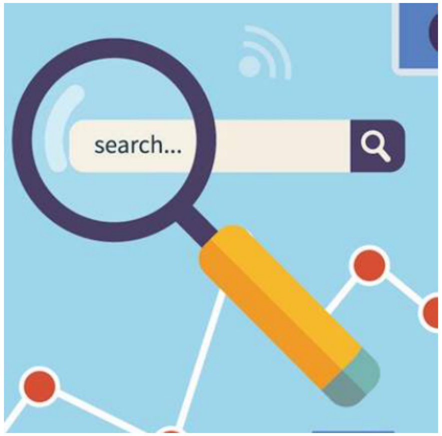
- Antiviral Agents / therapeutic use*
- Betacoronavirus
- Chloroquine / therapeutic use*
- Clinical Trials as Topic
- Coronavirus Infections / drug therapy*
- Humans
- Hydroxychloroquine / therapeutic use*
- Pandemics
- Pneumonia, Viral / drug therapy*

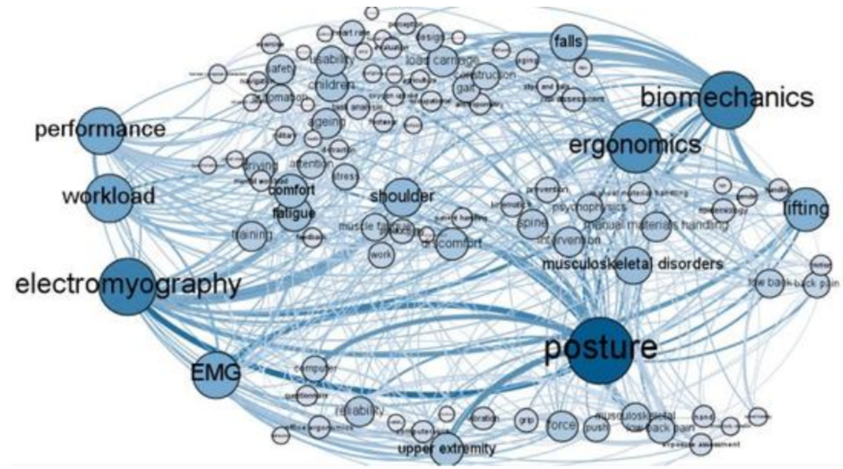
Figure 1. An Example of Biomedical Literature Semantic Indexing Problem.

What is it used for?

1. Document Categorization
2. Knowledge-base Construction
3. Search Engine
4. Drug Discovery
5. COVID-19 Q&A System
6. More…

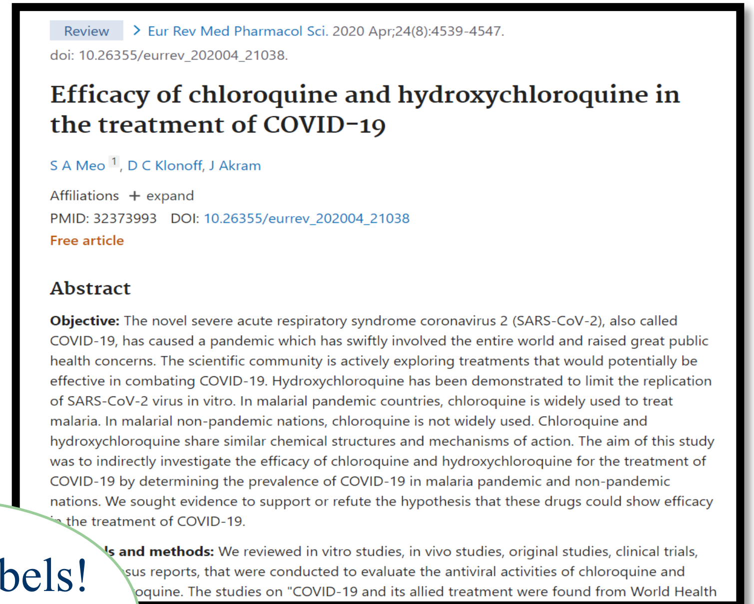






Extreme Multi-Label Classification (XMC)

- Large Label Set (extreme)
- Multi-Label Classification



Review > Eur Rev Med Pharmacol Sci. 2020 Apr;24(8):4539-4547.
doi: 10.26355/eurrev_202004_21038.

Efficacy of chloroquine and hydroxychloroquine in the treatment of COVID-19

S A Meo [1], D C Klonoff, J Akram

Affiliations + expand

PMID: 32373993 DOI: 10.26355/eurrev_202004_21038

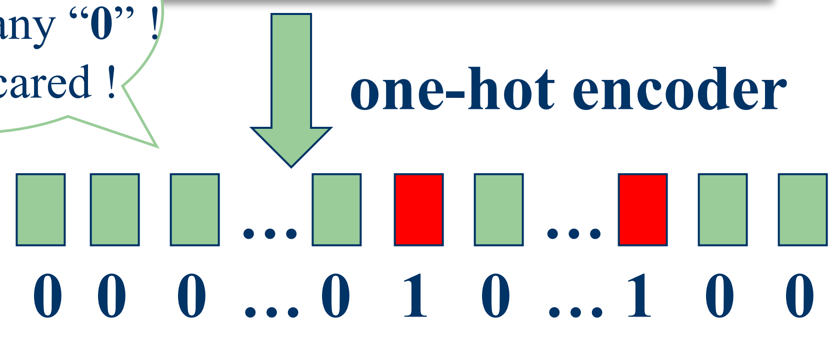
Free article

Abstract

Objective: The novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), also called COVID-19, has caused a pandemic which has swiftly involved the entire world and raised great public health concerns. The scientific community is actively exploring treatments that would potentially be effective in combating COVID-19. Hydroxychloroquine has been demonstrated to limit the replication of SARS-CoV-2 virus in vitro. In malarial pandemic countries, chloroquine is widely used to treat malaria. In malarial non-pandemic nations, chloroquine is not widely used. Chloroquine and hydroxychloroquine share similar chemical structures and mechanisms of action. The aim of this study was to indirectly investigate the efficacy of chloroquine and hydroxychloroquine for the treatment of COVID-19 by determining the prevalence of COVID-19 in malaria pandemic and non-pandemic nations. We sought evidence to support or refute the hypothesis that these drugs could show efficacy in the treatment of COVID-19.

...ls and methods: We reviewed in vitro studies, in vivo studies, original studies, clinical trials, ...sus reports, that were conducted to evaluate the antiviral activities of chloroquine and ...oquine. The studies on "COVID-19 and its allied treatment were found from World Health

>10k Labels!
Too Many "0" !
I am scared !



- Urgency
  - high cost
  - rapid increase of COVID-19 literature
  - effective and robust XMC technologies

Extreme Multi-Label Classification (XMC)

- One vs all models
  - Treat each label as an independent binary classification problem. (e.g. Rohit et al., 2017; Ian et al., 2017)
- Embedding based models
  - Represent target labels in a low-dimensional embedding space. (e.g. Kush et al., 2015; Yukihiro et al., 2017)
- Tree based models
  - Reduce the computational cost by statistics features to create a tree hierarchy for labels. (e.g. Himanshu et al., 2016; Kalina et al., 2016; Sujay et al., 2019;)
- Deep learning models
  - Leverage deep neural networks to encode and represent document text. (e.g. Xun et al., 2016; You et al., 2018; Jin et al., 2018; Chang et al, 2019; Xun et al., 2020;)

COVID-19 Semantic Indexing Corpus (CSIC)

- Raw Data
  - The COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020)
    - Large data size (59k documents)
    - Considerable coverage (3k journals, over 92% in Biology, Medicine, and Chemistry)
    - Abundant meta information (19 property fields, e.g. text, ID, sha, license, data sources)
    - Widely used for various COVID-19 research and competition (Kaggle Challenge, TREC-COVID Challenge, etc.)
    - Doesn't have any semantic labels
  - MedLine database
  - PubMed / PMC Central

COVID-19 Semantic Indexing Corpus (CSIC)

- The CSIC Corpus Construction
  a) Metadata extraction from CORD19 corpus
  b) Metadata extraction from MedLine database
  c) Webpage collection & metadata extraction from PubMed / PMC Central
  d) Metadata & label mapping from different resources
  e) Redundant and conflict documents integration

CSIC Corpus

- More than 20 different informative fields are retained in CSIC corpus
  - title, abstract, body text, keywords, author, affiliation, journal, date, data source, etc.

Table 1. The Statistic Information of The CSIC Corpus.

Dataset	#Documents	#Titles	#Abstracts	#Body Texts	#Tokens	#Semantic Labels
CSIC	84,253	84,253	70,477	46,882	188,559,895	18,476

where ~10% of the CSIC corpus was reserved as the test set, i.e. the training set and the test set were 76,253 and 8000, respectively.

Table 2. The Distribution of the MeSH Semantic Topics in The CSIC Corpus.

MeSH Semantic Topics	% of Terms
Diseases	26.08
Analytical Diagnostic and Therapeutic Techniques and Equipment	14.74
Chemicals and Drugs	13.95
Health Care	12.92
Organisms	10.39
Phenomena and Processes	6.72
Anatomy	5.22
Named Groups	3.60
Geographicals	1.31
Information Science	1.24
Disciplines and Occupations	1.24
Anthropology Education Sociology and Social Phenomena	0.94
Psychiatry and Psychology	0.94
Technology Industry and Agriculture	0.57
Humanities	0.14

Correlation Neural Networks (Xun et al., 2020)

Take advantage of the correlations among different target labels by correlation units.

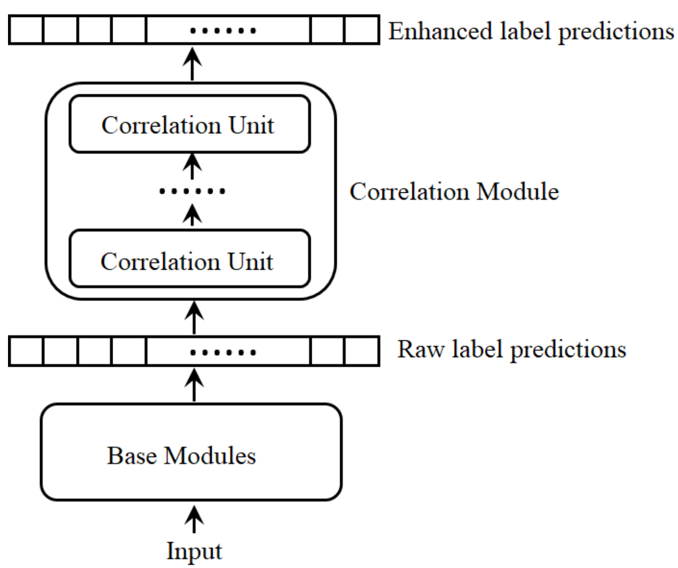


Figure 2. The Structure of A Correlation Network.

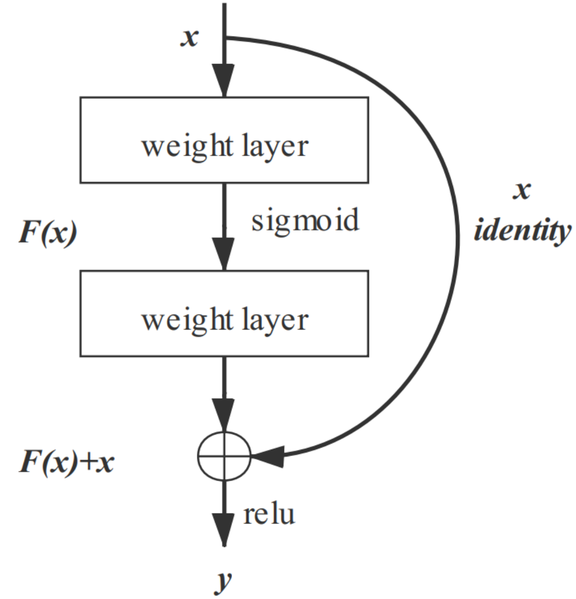


Figure 3. The Structure of A Correlation Unit.

where the base modules consist of four SOTA systems, i.e. **BertXML**(Chang et al., 2019), **XMLCNN** (Liu et al.,2017), **MeSHProbeNet** (Xun et al., 2016) and **AttentionXML** (You et al., 2018).

Instance-based Evaluation Metrics

- Precision at top k (P@K)
- Normalized Discounted Cumulative Gain at top k (N@K).

$$\text{precision@k} = \frac{1}{k} \sum_{l \in r_k(\hat{z})} z_l,$$

$$\text{DCG@k} = \sum_{l \in r_k(\hat{z})} \frac{z_l}{\log(l+1)},$$

$$\text{nDCG@k} = \frac{\text{DCG@k}}{\sum_{l=1}^{\min(k, \|z\|_0)} \frac{1}{\log(l+1)}},$$

- where $\boldsymbol{z} \in \{0, 1\}^L$ denote the ground truth label vector of an instance; $\boldsymbol{z}\hat{} \in \mathrm{R}^L$ denote the model predicted score and $r_k(\boldsymbol{z}\hat{})$ is the ground truth indices corresponding to the top k indices of the model predicted rank list

Experimental results

- Experimental Settings:
  - Evaluation: P@1, P@3, P@5, N@1, N@3, N@5
  - Number of correlation units: 2

Table 3. Performance Comparison of Different Systems.

Model	P@1	P@3	P@5	N@1	N@3	N@5	#GPUs	#Hours	model size(GB)
XMLCNN	93.66	79.86	70.80	93.66	82.98	76.12	1	0.50	0.38
Correl-XMLCNN	94.22	81.48	72.60	94.22	84.53	77.67	1	0.62	0.65
BertXML	94.12	83.01	73.50	94.12	85.57	78.58	1	1.25	0.45
Correl-BertXML	93.36	83.99	75.04	93.36	86.23	79.72	1	1.58	0.73
MeSHProbeNet	94.80	83.09	74.33	94.80	85.79	79.28	1	1.60	0.55
Correl-MeSHProbeNet	94.82	84.63	75.72	94.82	87.11	80.57	1	2.17	0.83
AttentionXML	94.32	85.62	**78.18**	94.32	87.82	82.17	4	8.17	0.37
Correl-AttentionXML	**94.82**	**85.70**	77.91	**94.82**	**87.84**	**82.17**	4	8.50	0.65

Figure 4. Training curves on the validation set of the CSIC corpus.

- Dotted lines denote basic models and solid lines denote target correlation based models.

Conclusion

- Correlation networks are able to consistently improve the performance of the existing XMC models;
- The deeper the mode is, the less the improvements with correlation networks;
- Among all the deep models, the improvement over XMLCNN is the largest, the improvement over AttentionXML is the smallest;
- Correlation networks exhibits the ability to accelerate the convergence rate during the training process;

Reference

- Rohit Babbar and Bernhard Schölkopf. 2017. Dismec: Distributed sparse machines for extreme multi-label classification. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. 721–729.
- Ian EH Yen, Xiangru Huang, Wei Dai, Pradeep Ravikumar, Inderjit Dhillon, and Eric Xing. 2017. Ppdsparse: A parallel primal-dual sparse method for extreme classification. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 545–553.
- Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme multilabel loss functions for recommendation, tagging, ranking & other missing label applications. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 935–944.
- Kalina Jasinska, Krzysztof Dembczynski, Róbert Busa-Fekete, Karlson Pfannschmidt, Timo Klerx, and Eyke Hullermeier. 2016. Extreme F-measure maximization using sparse probability estimates. In International Conference on Machine Learning. 1435–1444.
- Sujay Khandagale, Han Xiao, and Rohit Babbar. 2019. Bonsai-diverse and shallow trees for extreme multi-label classification. arXiv preprint arXiv:1904.08249 (2019).
- Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. 2015. Sparse local embeddings for extreme multi-label classification. In Advances in neural information processing systems. 730–738.
- Yukihiro Tagami. 2017. Annexml: Approximate nearest neighbor search for extreme multi-label classification. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 455-464.
- Guangxu Xun, Kishlay Jha, Ye Yuan, Yaqing Wang, and Aidong Zhang. 2016. MeSHProbeNet: A Self-attentive Probe Net for MeSH Indexing. Bioinformatics 32, 12 (2016), 70–79.
- Dai S, You R, Lu Z, et al. FullMeSH: improving large-scale MeSH indexing with full text[J]. Bioinformatics, 2020, 36(5): 1533-1541.
- Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Reference

- Ronghui You, Suyang Dai, Zihan Zhang, Hiroshi Mamitsuka, and Shanfeng Zhu. 2018. Attentionxml: Extreme multi-label text classification with multilabel attention based recurrent neural networks.
- Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2018. AttentionMeSH: Simple, Effective and Interpretable Automatic MeSH Indexer. In Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering. 47–56.
- Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit Dhillon. 2019. X-BERT: eXtreme Multi-label Text Classification with BERT. arXiv preprint arXiv:1905.02331 (2019).
- Guangxu Xun, Kishlay Jha, Jianhui Sun, et al. 2020. Correlation Networks for Extreme Multi-label Text Classification. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1074-1082.
- Wang L L, Lo K, Chandrasekhar Y, et al. CORD-19: The Covid-19 Open Research Dataset. arXiv preprint arXiv:2004.10706, 2020.
- Mao Y, Lu Z. MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank[J]. Journal of biomedical semantics, 2017, 8(1): 15.
- Dai S, You R, Lu Z, et al. FullMeSH: improving large-scale MeSH indexing with full text[J]. Bioinformatics, 2020, 36(5): 1533-1541.
- Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit Dhillon. 2019. X-BERT: eXtreme Multi-label Text Classification with BERT. arXiv preprint arXiv:1905.02331 (2019).
- Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 115–124.

Thanks!