The Impact of Interdisciplinarity and Entity Characteristics on the Clinical Translation Intensity of COVID-19 Papers

Shuang Chen[1], Chunli Liu[2, *]

[1] *School of Health Management, China Medical University, Shenyang, China, 110122*
[2] *Library, China Medical University,Shenyang, China, 110122*

Abstract

Biomedical research has not only academic impact, but also clinical impact. The evaluation of the clinical impact of COVID-19 papers is very interesting. It is still unclear whether the interdisciplinary and entity characteristics of the paper affect the clinical impact. We selected COVID-19 papers published in 2021 for preliminary exploration and got some interesting findings. We found that only 22.43% were cited in clinical trials or clinical guidelines; 47.42% of the papers are biased towards human research in MeSH terminology; On average, 46.1% of papers have the potential to be cited in clinical studies after publication. The interdisciplinary features of the paper are not significantly related to clinical translation intensity, but the biomedical entity features mentioned in the paper are significantly related to clinical translation intensity. The number of Chemical entities, Gene entities and Species entities had significant negative effects on clinical translation intensity. However, the number of Disease entities mentioned in the paper has positive impact on the clinical translation intensity.

Keywords

Interdisciplinarity, Entity, Characteristics, Clinical Translation Intensity, COVID-19

1 Introduction

The main purpose of biomedical research is to serve the public health and improve the well-being of the people. So, Biomedical research should have not only academic impact but also clinical impact. The evaluation of scientific impact has been very rich, but the identification and measurement of clinical impact is still relatively weak. Interdisciplinary research is the main mode of modern biomedical research. Previous studies have shown that interdisciplinarity is significantly related to the citation of papers, and highly cited papers have a higher interdisciplinary level. In our previous research on papers published by Lasker Prize winners in Basic Medicine, we found that the average number of disciplinary involved in papers published by winners was positively correlated with APT, while the number of subjects is not significantly correlated with clinical citations [1].

Therefore, more diverse, and complex interdisciplinary indicators are needed to verify its correlation with clinical translation intensity in other samples. Recently, biological entity features have also been used to predict the clinical translational potential of a paper or to measure the translational progress of a paper [2-3]. However, it is still unknown which biological entity characteristics affect the clinical translation intensity of papers.

This study includes two objectives: 1) to measure the clinical translation intensity of COVID-19 articles published in 2021; 2) to test the impact of interdisciplinary level and the

Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023) and the 3rd AI + Informetrics (AII2023), June 26, 2023, Santa Fe, New Mexico, USA and online
EMAIL: liuchunliliangxu@163.com (Chunli Liu)

characteristics of biological entity on the intensity of clinical translation of COVID-19 papers.

2 Related works

2.1 The measurement of clinical translation intensity of biomedical papers

As for the clinical translation intensity of papers, some methods have been developed, such as “Research Level” [4], “Biomedical Triangle method” [5], “Level Score” [6], and clinical citations [7-9]. Hutchins et al. developed the "Approximate potential to translate scores (APT)" based on the "biomedical triangle" method combined with the random forest model [10]. “Translational Progression (TP)” has been recently proposed from the perspective of biomedical entities to track the clinical translation intensity of biomedical papers [3].

2.2 Interdisciplinary features of COVID-19 papers

It is reported that in 2020, the research on COVID-19 involved an average of 6-7 disciplines [11]. Liu's research confirmed that the global coronavirus pandemic was a "catalyst" for scientific innovation [12]. Zhang et al. investigated the interdisciplinary of COVID-19 papers published in 2020, the first year after the outbreak of COVID-19 [11]. In recent years, many studies have confirmed that there is a certain relationship between interdisciplinary research and the citation impact of papers [1-15]. However, in contrast, our recent study found no significant correlation between the number of disciplines and the clinical translation of papers [1].

2.3 Entity characteristics in biomedical papers

Knowledge entity refers to the unit of knowledge in scientific articles [16]. The biomedical entities in biomedical articles mainly include diseases, genes, drugs, pathways, and CellLine [17]. Li X. has developed four indicators (Popularity, Promising, Prestige and collaboration) based on the characteristics of biomedical entities. [17]. Li X. also predicted the clinical translation potential of scientific papers based on the entity characteristics of scientific papers [2-3].

The entity characteristics of a paper are related to the interdisciplinary characteristics, for example, the entity of a biomedical paper is a terminology or a unit of knowledge in the biomedical field. But the two are measured from different perspectives. The interdisciplinary features are measured from the subject of the journal in which the paper's references are published, while biomedical entity features are measured from the biomedical proprietary concepts mentioned in the title and abstract. Entity features are more fine-grained than subject features.

3 Methods

The research process mainly includes three parts: First, data collection and preprocessing; The second is to measure the clinical translation intensity of papers; Third, verify the effect of interdisciplinarity and entity characteristics on the clinical translation intensity of papers.

3.1 Data collection

On March 13, 2023, we exported a total of 120,573 papers published in 2021 from the iSearch COVID-19 portfolio [18]. The iSearch COVID-19 Portfolio tool (https://icite.od.nih.gov/covid19/search/) was developed and implemented by the Office of Portfolio Analysis (OPA) of NIH. According to the requirements of previous scholars for calculating interdisciplinary indicators and extracting biomedical entities, we adopted a series of

inclusion and exclusion criteria, as shown in Figure 1. Finally, we obtained a sample consisting of 36,797 COVID-19 papers published in 2021.

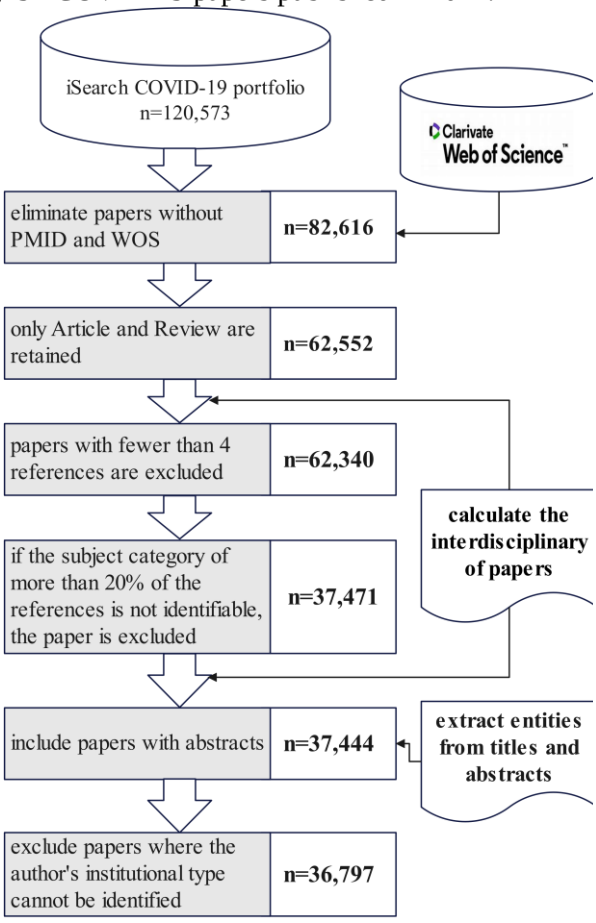

Figure 1: Inclusion and exclusion criteria

3.2 Variables

3.2.1 Dependent variables

We select three indicators (APT, Human and Cited_by_Clin) to represent the clinical translation intensity of the paper from different perspectives. Data for all three dependent variables are downloaded from the iCite platform [10]. Table 1 lists the three dependent variables and their implications. The iCite database is a retrieval platform developed by NIH, which provides the function of downloading the citation impact of papers and translation impact indicators through the PMID of papers.

APT (Approximate Potential to Translate) is a clinical translation indicator developed by Hutchins et al. in 2019 to predict the probability that a paper will be cited in a clinical paper shortly after publication [10]. The value of APT ranges from 0 to 1. The greater APT is, the higher the probability of the paper being cited by clinical papers after publication.

The indicator "Human" is another clinical translation indicator developed by Hutchins et al. which indirectly reflects whether the research topic of the paper focuses on human health. The range of "Human" value is between 0 and 1. The larger the "Human" value is, the closer the topic of the paper is to human health, and the higher the clinical translation potential of the paper is.

The indicator Cited_by_Clin refers to the absolute number of times each paper is cited by clinical papers such as clinical trials and clinical guidelines. Cited_by_Clin is a continuous variable that is greater than or equal to zero.

3.2.2 Independent variables

We select two kinds of independent variable indicators, including interdisciplinary characteristics and entity characteristics.

We measure interdisciplinary level by the diversity of subjects a paper is classified into. In this paper, we used Variety, 1-GINI, Disparity and Rao-Stirling indicator to measure the interdisciplinary level of the papers. Table 2 lists the interdisciplinary indicators and their connotations.

In this study, PubTator Central (PTC) was used to extract the biological entities mentioned in the title and abstract of each article (https://www.ncbi.nlm.nih.gov/research/pubtator/). It divides biological entities into six categories: Gene, Chemical, Disease, CellLine, Mutations, and Species. The efficiency and accuracy of PubTator

for biomedical entity extraction are superior to manual text mining [19].

In this study, Pubtator API was invoked to extract biomedical entities mentioned in the title and abstract of the paper in batches by PMID number. We assign a unique identifier to each individual entity of each entity class. The six indicators that separately count the number of entities of each type mentioned in each paper including CellLine, Chemical, Disease, Gene, Mutation and Species. We performed de-processing when calculating the number of entities mentioned in each paper. For example, if a paper mentions the same independent entity multiple times, we only count it once. Table 3 summarizes entity characteristic indicators and their definitions.

3.2.3 Covariates

We select 8 variables as control variables, which are 1) Whether it is a clinical study, clinical guidelines or clinical trial, extracted from iCite; 2) Paper length, extracted from the "PG" field of the core collection of WOS database; 3) Title length, counting the number of words in the title; 4) The number of references, extracted from the "NREF" field of the core collection of WOS database; 5) The number of authors, extracted from the "AU" field of the core collection of WOS database; 6) The number of funds, extracted from the "FU" field of the core collection of WOS database; 7) Paper language, extracted from the "LA" field of the core collection of WOS database. 8) The first corresponding author's institution type, extracted from the "RP" field of the core collection of WOS database.

3.3 Statistics Analysis

First, we calculate APT, Human, and Cited_by_Clin for COVID-19 papers published in 2021. In addition, we use the negative binomial regression method to examine the influence of interdisciplinary features and entity features on APT, Human and Cited_by_Clin of the papers.

4 Results

4.1 Summary of clinical translation intensity

The study sample was 36,797 papers. APT and Human have values between 0 and 1, while Cited_by_Clin has a maximum value of 168. Therefore, we use two vertical axes in box plot (Figure 2). APT and Human correspond to the scale on the left and Cited_by_Clin to the scale on the right. Table 1 lists the descriptive analysis result of dependent and independent variables.

Table 1
Descriptive analysis of dependent and independent variables

	Mean	Median	Min	Max
APT	0.461	0.5	0.05	0.95
Cited_by_Clin	0.495	0	0	168
Human	0.581	0.5	0	1
RAO-STIRLING	0.832	0.851	0	0.959
1-GINI	0.000	0	0	0.003
VARIETY	0.001	0	0	0.005
DISPARITY	0.909	0.919	0	0.991
CellLine	0.060	0	0	6
Chemical	0.652	0	0	22
Disease	3.536	3	0	41
Gene	0.638	0	0	27
Mutation	0.044	0	0	25
Species	1.558	1	0	19

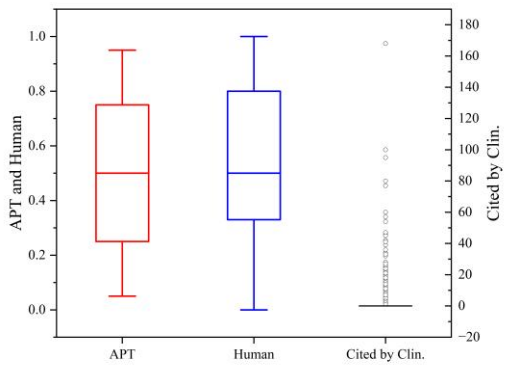

Figure 2: Box plot of APT, Human and Cited_by_Clin

As is shown in Figure 2, the median of Human is smaller than its average, suggesting that most papers with a larger Human value in the sample. 47.42% of the papers had a Human value greater than 0.5, and 77.57% were not cited in clinical papers. The distribution of APT is close to symmetric, while Human is positively skewed. Cited_by_Clin is a variable with most zeros, and its mean, median, upper, and lower quartiles and minimum values are all zero. We do not see IQR and whiskers, but only many positive outliers.

4.2 The impact of interdisciplinarity and entity features on the clinical translation intensity

Table 2

Negative binomial regression for APT

APT	Coef.	St.Err.	p-value
RAO-STIRLING	.084	.207	.687
VARIETY	-17.907	98.617	.856
1-GINI	30.889	146.358	.833
DISPARITY	-.052	.318	.87
Number of CellLine entities	-.025	.029	.395
Number of Chemical entities	-.038	.006	0
Number of Disease entities	.029	.003	0
Number of Gene entities	-.017	.006	.004
Number of Mutation entities	-.031	.02	.12
Number of Species entities	-.077	.008	0
Length of paper	.007	.002	0
Number of references	-.0004349	.000214	.042
Number of funds	.015	.005	.002
Length of Title	.005	.002	.004
Number of authors	.0003123	.000092	.001
Clinical Article	.228	.029	0
English Language	.917	.131	0
Type of institution			
Company (REF)	0	.	.
Foundation	.101	.111	.363
Government	.162	.107	.132
Hospital	.029	.067	.668
Research institution	-.091	.069	.192
University	-.037	.065	.57
Constant	-1.799	.241	0

Table 2 reveals the negative binomial regression results for APT. We found that the interdisciplinarity had no statistical significance on APT ($p > 0.1$). The number of chemical entities, gene entities and species entities were significantly

negatively correlated with APT, while the number of disease entities was significantly positively correlated with APT.

Table 3

Negative binomial regression for Human

Human	Coef.	St.Err.	p-value
RAO-STIRLING	.034	.18	.852
VARIETY	-89.1	89.152	.318
1-GINI	114.387	131.733	.385
DISPARITY	-.004	.276	.988
Number of CellLine entities	-.246	.034	0
Number of Chemical entities	-.052	.006	0
Number of Disease entities	.03	.002	0
Number of Gene entities	-.041	.006	0
Number of Mutation entities	-.054	.02	.009
Number of Species entities	-.122	.007	0
Length of paper	-.005	.002	.007
Number of references	-.002	.0002245	0
Number of funds	-.031	.005	0
Length of Title	.009	.001	0
Number of authors	.0001628	.0001227	.184
Clinical Article	.184	.026	0
English Language	-.019	.07	.788
Type of institution			
Company (REF)	0	.	.
Foundation	.169	.102	.098
Government	.168	.102	.099
Hospital	.174	.064	.006
Research institution	.021	.067	.749
University	.102	.063	.103
Constant	-.493	.191	.01

Table 3 shows that the interdisciplinarity has no significant impact on the Human of the papers. However, the number of CellLine entities, chemical entities, gene entities, mutation entities and species entities have a negative impact on Human, while the number of disease entities has a positive impact on Human.

Table 4

Negative binomial regression for Cited_by_Clin

Cited_by_Clin	Coef.	St.Err.	p-value
RAO-STIRLING	-.096	.353	.786
VARIETY	-28.332	167.529	.866
GINI	91.928	249.526	.713
DISPARITY	-.252	.536	.638
Number of CellLine entities	.264	.045	0
Number of Chemical entities	-.04	.01	0
Number of Disease entities	.066	.005	0

Number of Gene entities	-.026	.009	.005
Number of Mutation entities	-.15	.04	0
Number of Species entities	-.1	.014	0
Length of paper	.006	.003	.028
Number of references	.001	.0003638	.074
Number of funds	.072	.008	0
Length of Title	-.001	.003	.655
Number of authors	.039	.002	0
Clinical Article	1.095	.048	0
English Language	1.89	.257	0
Type of institution			
Company (REF)	0	.	.
Foundation	-.425	.193	.027
Government	.544	.171	.002
Hospital	-.128	.106	.227
Research institution	-.306	.11	.005
University	-.454	.103	0
Constant	-2.824	.425	0

As shown in Table 4, interdisciplinarity had no significant effect on the number of clinical citations. The number of chemical entities, gene entities, mutation entities and species entities have a significant negative impact on the number of clinical citations, while the number of CellLine entities and disease entities has a significant positive impact on the number of clinical citations.

5 Discussion

Firstly, after measuring the clinical translation intensity of 36,797 COVID-19 papers published in 2021, We found that APT, Human and Cited_by_Clin, three clinical translation intensity indicators with different connotations, can provide three dimensions of mutually complementary information for the same paper sample. APT is measured from the perspective of the potential of a paper to be cited by clinical guidelines or clinical trials after publication, Human is measured from the perspective of the proportion of MeSH terms in a paper that are biased toward human beings, and Cited_by_Clin is measured from the perspective of the actual number of citations of a paper by clinical guidelines or clinical trials.

Our team previously investigated the clinical conversion intensity of papers published by Lasker Prize winners in Basic Medicine, and found that the average value of APT and Cited_by_Clin was 0.24 and 0.59, respectively, and 80% of the papers were not cited in clinical papers [1]. By comparison, the average APT for COVID-19 papers was higher than the average for papers published by recipients of the Basic Medicine Prize. This may be because the topic of COVID-19 is more clinical, and the clinical translational potential of clinical papers is obviously higher than that of basic research. The number of clinical citations is lower than that of the papers of Basic Medicine Award winners, which may be since our paper collection only has one year of cross-sectional data, and the citation window is only two years, so the accumulated clinical citations will be affected by this factor.

Secondly, although most previous studies have confirmed that interdisciplinarity has a significant impact on the citations of papers, in this study, we did not detect a significant impact of interdisciplinarity on the clinical impact of COVID-19 papers. Some studies have found that although interdisciplinary research dominates the academic cooperation network, this competitive advantage does not translate into immediate returns, and the impact of these studies is low in the short term. Interdisciplinary research requires more time and perseverance to overcome challenges [20].

We believe that interdisciplinary research is more time consuming and energy consuming, which will not lead to higher impact performance

in the short term. In addition, clinical impact puts more emphasis on the flow of knowledge in a paper to clinical application, while the traditional method of measuring interdisciplinary level is very crude, representing the discipline of the paper from the discipline of the journal in which the reference was published. This may lead to the traditional interdisciplinary level is not correlated with the clinical translation strength of the paper.

In contrast, a significant relationship was examined between the entity characteristics of the paper and its clinical impact. The number of chemical entities, gene entities and species entities had significant negative effects on APT. The number of CellLine entities, chemical entities, gene entities, species entities and mutation entities had significant negative effects on Human. The number of chemical entities, gene entities, mutation entities and species entities had significant negative effects on clinical citation. The CellLine entities has a positive relationship with clinical citation. Interestingly, the number of disease entities had a positive impact on APT, Human, and clinical citations.

Some chemical entities are often mentioned in COVID-19 papers. For example, the effectiveness of methylprednisolone in treating high-risk COVID-19 patients [21] and the efficacy and safety of tocilizumab in the treatment of severe COVID-19 patients by blocking IL-6 to suppress inflammatory cytokine storm immune response [22]. These studies are focused on the efficacy of a particular drug to treat COVID-19 patients, which is groundbreaking and innovative research, and naturally, the clinical translation intensity of this type of research is higher.

Similarly, when genome researchers explore the relationship between COVID-19 and host genes, genes related to COVID-19 critical illness, and genomics studies on the natural origin of COVID-19, the fewer the number of genetic entities mentioned in the paper may indicate the stronger the breakthrough of the paper content and the higher the clinical translation intensity.

For species entities, the lower the number of species entities mentioned in the paper, the higher the clinical translation intensity. For example, specifically studying the clinical presentation, pathogenesis, or treatment of COVID-19 in only men or women, or in only children or the elderly, will likely be more valuable for clinical diagnosis, treatment, and prevention.

Studies of the effect of drugs or vaccines targeting a single protein, DNA, and SNP mutation on the treatment or prevention of COVID-19 will have higher clinical translational strength than studies that combine multiple mutations.

Cytology studies on COVID-19 also follow a rule, that is, the CellLine entities involved in the research are more specific, the research is more in-depth, and the clinical application is more valuable for reference, so the APT and Human of the paper is higher.

However, if there are more disease entities mentioned in COVID-19 papers, it means that this is a study on severe COVID-19 patients with multiple infections or comorbidities, which is a breakthrough study on severe COVID-19 patients, so this kind of research has higher clinical translation intensity.

The paper has some limitations. First, the intensity of clinical translation of a paper varies with the time of publication. COVID-19 papers published in 2021 have only a window of nearly two years to accumulate citations. Second, only some simple entity characteristic variables were used in this study, and future work is to create more diversified and systematic entity characteristic variables to further verify their relationship with clinical translation intensity.

6 Conclusion

We measured the clinical translation strength of COVID-19 papers published in 2021 and found that only 22.43% were cited in clinical trials or clinical guidelines; 47.42% of the papers are biased towards human research in MeSH terminology; On average, 46.1% of papers have the potential to be cited in clinical studies after publication. The interdisciplinary features of the paper are not

significantly related to clinical translation, but the biomedical entity features mentioned in the paper are significantly related to clinical translation. The more disease entities mentioned in the paper, the stronger the clinical translation; However, the more chemical, genetic and species entities are mentioned, the weaker the clinical translation intensity is.

References

[1] Zang D., Liu C. (2023). Exploring the clinical translation intensity of papers published by the world's top scientists in basic medicine. Scientometrics 4, 2371-2416.

[2] Li X., Tang X., Cheng Q. (2022). Predicting the clinical citation count of biomedical papers using multilayer perceptron neural network. Journal of Informetrics 4, 101333.

[3] Li X., Tang X., Lu W. (2023). Tracking biomedical articles along the translational continuum: a measure based on biomedical knowledge representation. Scientometrics 2, 1295–1319.

[4] Narin F., Pinski G., Gee H. (1976). Structure of the Biomedical Literature. Journal of the American Society for Information Science 1, 25-45.

[5] Weber G. M. (2013). Identifying translational science within the triangle of biomedicine. Journal of translational medicine 11, 126.

[6] Ke Q. (2019). Identifying translational science through embeddings of controlled vocabularies. Journal of the American Medical Informatics Association 6, 516-523.

[7] Thelwall M., Maflflahi N. (2016). Guideline references and academic citations as evidence of the clinical value of health research: Guideline References and Academic Citations as Evidence of the Clinical Value of Health Research. Journal of the Association for Information Science and Technology 4, 960–966.

[8] Lewison G., Sullivan R. (2008). The impact of cancer research: How publications influence UK cancer clinical guidelines. British Journal of Cancer 12,1944-1950.

[9] Grant J., Cottrell R., Cluzeau F., Fawcett G. (2000). Evaluating "payback" on biomedical research from papers cited in clinical guidelines: Applied bibliometric study. BMJ 320,1107-1111.

[10] Hutchins B. I., Davis M.T., Meseroll R. A., Santangelo G. M. (2019). Predicting translational progress in biomedical research. PLoS biology 10, e3000416.

[11] Zhao Y., Liu L., Zhang C. (2022). Is coronavirus-related research becoming more interdisciplinary? A perspective of co-occurrence analysis and diversity measure of scientific articles. Technological Forecasting and Social Change 175, 121344.

[12] Liu M., Bu Y., Chen C., Xu J., Li D., Leng Y., Freeman R. B., Meyer E. T., Yoon W., Sung M., Jeong M., Lee J., Kang J., Min C., Song M., Zhai Y., Ding Y. (2021). Pandemics are catalysts of scientific novelty: Evidence from COVID-19. Journal of the Association for Information Science and Technology 8, 1065-1078.

[13] Leydesdorff L., Wagner C. S., Bornmann L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. Journal of Informetrics 1, 255–269.

[14] Chen S., Arsenault C., Larivière V. (2015). Are top-cited papers more interdisciplinary? Journal of Informetrics 4,1034–1046.

[15] Chen S. J., Song Y. H., Shu F., Lariviere V. (2022). Interdisciplinarity and impact: the effects of the citation time window. Scientometrics 127, 2621–2642.

[16] Ding Y., Song M., Han J., Yu Q., Yan E., Lin L., Chambers T. (2013). Entitymetrics: measuring the impact of entities. PLoS ONE 8, e71416.

[17] Li X., Rousseau J. F., Ding Y., Song M., Lu W. (2020). Understanding Drug Repurposing from the Perspective of Biomedical Entities and Their Evolution: Bibliographic Research Using Aspirin. JMIR medical informatics 6, e16739.

[18] Scheinfeld L. (2022). LitCovid, iSearch COVID-19 portfolio, and COVID-19 Global literature on coronavirus disease. Journal of the Medical Library Association 2, 279-280.
[19] C Wei.H., Allot A., Leaman R., Lu Z.Y. (2019). PubTator Central: Automated Concept Annotation for Biomedical Full Text Articles. Nucleic Acids Res 47, w587-w593.
[20] Sun Y., Livan G., Ma A., Latora V. (2021). Interdisciplinary researchers attain better long-term funding performance. Commun Phys 4, 263.
[21] Gong Y., Guan L., Jin Z., Chen S.X., Xiang G. M., Gao B. (2020). Effects of methylprednisolone use on viral genomic nucleic acid negative conversion and CT imaging lesion absorption in COVID patients under 50 years old. Journal of Medical Virology 11, 2551-2555.
[22] Campochiaro C., Della-Torre E., Cavalli G., Luca G. D., Ripa M., Nicola B., Tomelleri A., Baldissera E., Rovere-Querini P., Ruggeri A., Monti G., Cobelli F. D., Zangrillo A., Tresoldi M., Castagna A., Dagna L., T-R. Study Group. (2020). Efficacy, and safety of tocilizumab in severe COVID-19 patients: a single-centre retrospective cohort study. European Journal of Internal Medicine 76, 43-49.