

Pre2Pub—Tracking the Path From Preprint to Journal Article: Algorithm Development and Validation

Mozhdeh Hashemzadeh

PhD Candidate in Medical Librarianship and Information Science

Isfahan University of Medical Sciences

29/Tir/1401

journal of medical internet research: 2022;24(4)

Why this article?

- Article type: Original Paper
- ISSN: 1438-8871
- Indexed in: ISI, Scopus, PubMed,
- Publisher/Holder: JMIR Publications
- “Tracking methods ”



2020 Impact Factor: 5.43

Study's goals

- ❖ investigating the extent to which the relationship information between preprint and corresponding journal publication is present in the published metadata
- ❖ how it can be further completed
- ❖ how it can be used in preVIEW to identify already republished preprints and filter those duplicates in search results

Introduction

- The publication of preprints, has gained popularity in recent years. the current COVID-19 pandemic shed new light on this type of publication because it allows researchers to communicate new findings quickly,
- Although preprints can be a very valuable source of information, there is a wide range of quality that cannot be assessed without close examination of the content
- The WHO titles the overabundance of both correct and incorrect information during a disease outbreak an “infodemic”. Therefore, preprints must be carefully integrated into existing information infrastructures





The ZB MED preprint Viewer preVIEW: COVID-19 includes 44.780 COVID-19 related preprints from [medRxiv \(15.212\)](#) and [bioRxiv \(5.332\)](#), [ChemRxiv \(275\)](#), [ResearchSquare \(8.428\)](#), [arXiv \(5.961\)](#) and from [Preprints.org \(1.793\)](#). The service has last been updated 9 hours ago. We try to update the data on a daily basis (7am CET/CEST).

The web page can also be accessed via [API](#).

If you use our application, please cite: [10.3233/SHTI210124](#) and [10.32384/jeahil17484](#)

[Query builder](#) [Expert Search](#) [Help](#)

Multi

Q enter a search term

Reset

Q Search

Feedback

Publication Date



Refine your search. Currently 44.780 documents are matched

< 1 2 3 4 5 ... 2239 >

Documents per page: 20

SARS-CoV-2 infection -induced immunity and the duration of viral shedding: results from

a Nicaraguan household cohort study

enable feedback mode

Expand all abstracts

Highlight semantic

What is preVIEW?

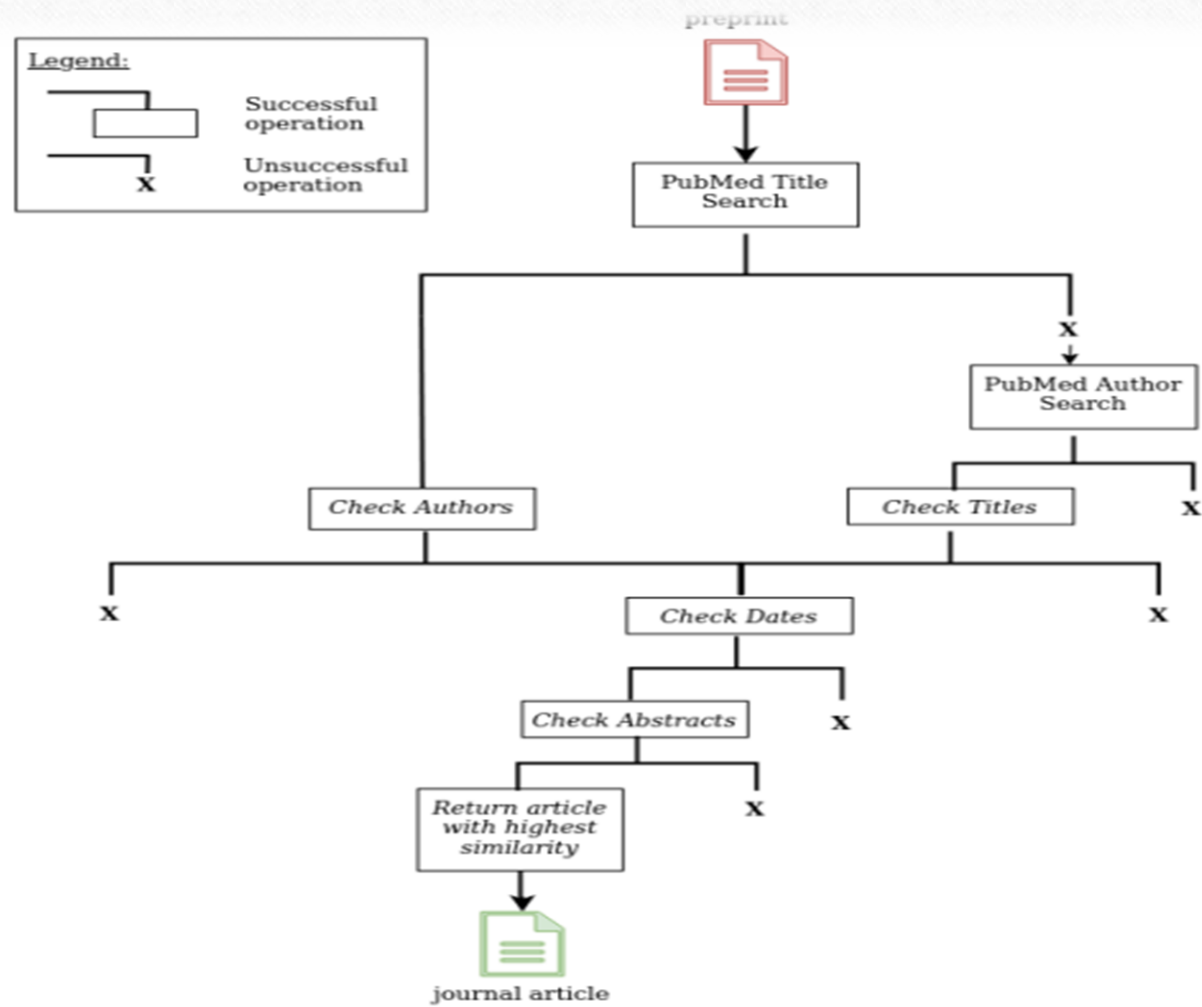
- a new preprint service and semantic search engine that currently combines more than 40,000 COVID-19–related preprints from 7 different sources
- Use Pre2Pub Algorithm
- Update data on a daily basis (7am)

What is Pre2Pub Algorithm?

- ✓ This algorithm uses a preprint DOI as input and searches for corresponding journal article in PubMed

- ✓ Has 5 steps:
 1. Compare titles
 2. Compare author's name after title review
 3. Search authors name in author field
 4. Compare the dated of publication
 5. Compare abstracts

Pre2Pub Algorithm



preVIEW method

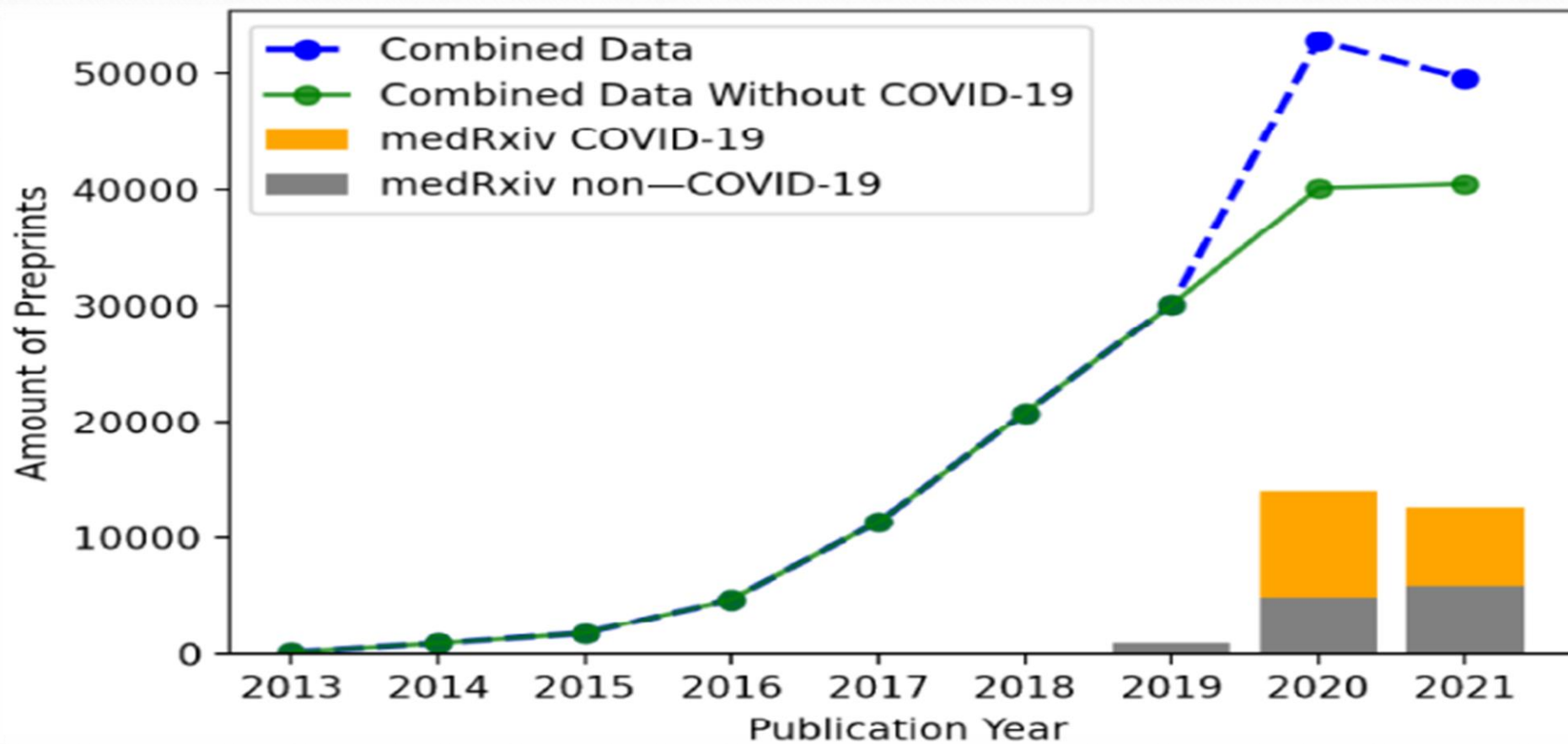
The screenshot shows the bioRxiv website homepage. At the top left is the CSH Cold Spring Harbor Laboratory logo and the bioRxiv logo with the tagline "THE PREPRINT SERVER FOR BIOLOGY". Navigation links include HOME, ABOUT, SUBMIT, NEWS & NOTES, ALERTS / RSS, and CHANNELS. A yellow box contains a disclaimer: "bioRxiv posts many COVID-19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive." Below this, it states "23,897 Articles (18,190 medRxiv, 5,707 bioRxiv)". The "Most recent first" section features an article titled "Epidemiological characteristics and transmission dynamics of the outbreak caused by the SARS-CoV-2 Omicron variant in Shanghai, China: a descriptive study" by Chen, Z., Deng, X., Fang, L., Sun, K., Wu, Y., Che, T., Zou, J., Cai, J., Liu, H., Wang, Y., Wang, T., Tian, Y., Zheng, N., Yan, X., Sun, R., Xu, X., Zhou, X., Ge, S., Liang, Y., Yi, L., Yang, J., Zhang, J., Ajelli, M., Yu, H., with a preprint ID of 10.1101/2022.06.11.22276273. A "Subject Areas" sidebar lists categories like Animal Behavior and Cognition, Biochemistry, Bioengineering, Bioinformatics, Biophysics, Cancer Biology, Cell Biology, Clinical Trials*, Developmental Biology, and Ecology. At the bottom, another article is partially visible: "How do Urban Factors Control the Severity of COVID-19?" by Roxon, J., Dumont, M.-S., Vilain, E., Pellenq, R., with a preprint ID of 10.1101/2022.06.17.22276576.

The screenshot shows the Crossref website homepage. At the top right is the Crossref logo. A navigation menu includes Apply, Members, Documentation, and Community for. A search bar is prominently displayed with the text "How can we help you?". Below the search bar, a quote reads: "Crossref makes research objects easy to find, cite, link, assess, and reuse. We're a not-for-profit membership organization that exists to make scholarly communications better." At the bottom, there are links for Board and Governance, Annual Meeting, and Blog. The background image is a sunset over a lake with a boat.

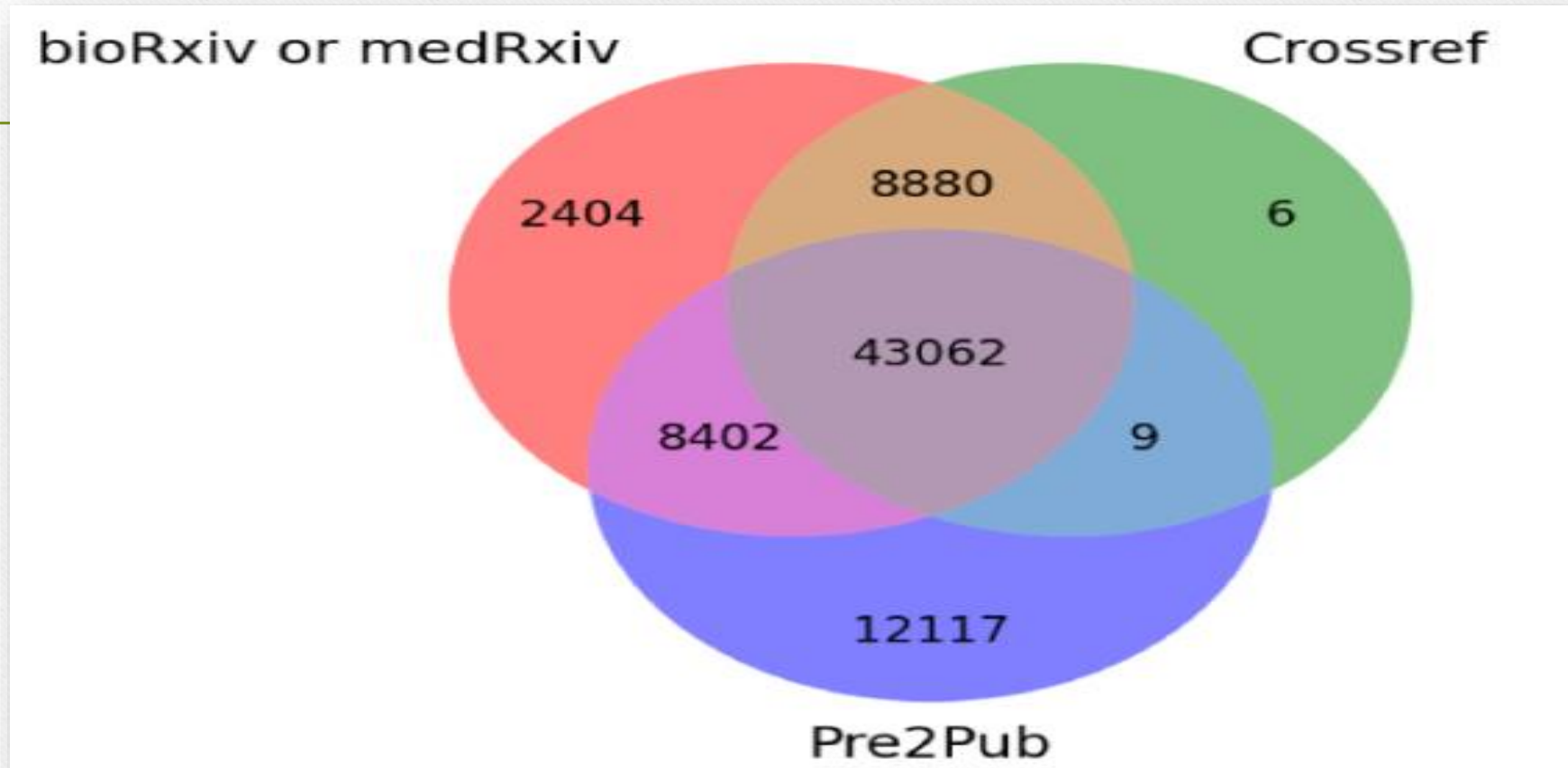
Results

The link from a preprint to its corresponding journal publication is not completely covered in the metadata of the preprint servers or in Crossref. The algorithm Pre2Pub is able to find approximately 16% more related journal articles with a precision of 99.27%. As long as there is no transparent and complete way to store this relationship in metadata, the Pre2Pub algorithm is a suitable extension to retrieve this information

Growth of preprint publications



Amount and intersection of found journal articles for 3 different methods



Evaluation of Pre2Pub

Table 2. Evaluation results of Pre2Pub on training and test data.

Data set	Precision, %	Recall, %	F ₁ -score, %
Training data	99.10	82.64	90.13
Test data	99.27	81.14	89.29

Table 3. Examples of false-positive and false-negative matches found by Pre2Pub

Preprint digital objective identifier	Journal digital object identifier found		Error Analysis
	bioRxiv/medRxiv	Pre2Pub	
10.1101/669713	10.1016/j.bpj.2019.11.1890	10.1016/j.bpj.2020.02.011	bioRxiv links only to conference abstracts, whereas Pre2Pub finds the corresponding article
10.1101/845933	10.1016/j.cub.2020.03.005	10.1111/ejn.15056	The correct article is not indexed in PubMed; the found article has the same first and last author and describes a related topic
10.1101/503763	10.1016/j.neuroimage.2019.1161	10.1016/j.neuroimage.2019.116186	Broken link at bioRxiv (incomplete digital object identifier); Pre2Pub finds the correct article
10.1101/549840	10.1128/mBio.00388-19	— ^a	No results via the application programming interface ^b
10.1101/2020.03.06.980631	10.1039/D0GC00903B	—	Article is not indexed in PubMed
10.1101/405597	10.1523/JNEUROSCI.0555-20.2020	—	Titles differ too much

Overview of the amount of bio- and medRxiv preprints

Year	COVID-19	Amount, n	Published, n (%)
2019	No	30,094	22,776 (75.68%)
2020	No	40,862	28,177 (68.96%)
2020	Yes	11,918	6920 (58.06%)
2021	No	40,441	13,489 (33.35%)
2021	Yes	9024	3518 (38.98%)

Thank you

