科学技術・学術政策研究所における最新の研究成果について

- 新型コロナウイルス感染症に関するプレプリントを用いた研究動向分析
  [DP-181, 2020年5月15日 公表]
  [DP-186, 2020年6月30日 公表, 2020年11月4日 補遺公表]

- 博士人材追跡調査 第3次報告書 [NR-188, 2020年11月27日 公表]

2021年1月

科学技術・学術政策研究所

NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY POLICY

新型コロナウイルス感染症に関するプレプリントを用いた研究動向分析

※ **新型コロナウイルス感染症** (COVID-19)
新型コロナウイルス (SARS-CoV-2)

本資料は、以下の報告書のポイントを示したものです。
・「COVID-19 / SARS-CoV-2 に関する研究の概況」, DISCUSSION PAPER-181, 文部科学省科学技術・学術政策研究所, 2020年5月15日公表.
・「COVID-19 / SARS-CoV-2 関連のプレプリントを用いた研究動向の試行的分析」, DISCUSSION PAPER-186, 文部科学省科学技術・学術政策研究所, 2020年6月4日公表、2020年11月4日補遺公表.

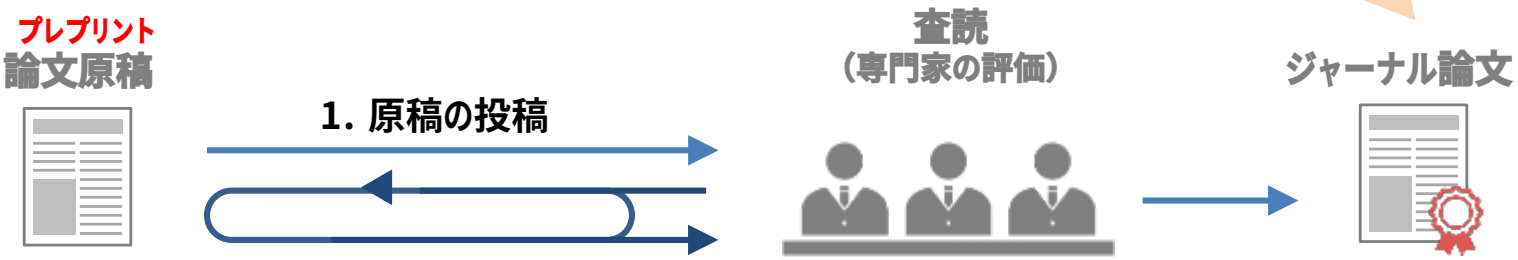
DOI: http://doi.org/10.15108/dp181
http://doi.org/10.15108/dp186

プレプリントとは？

- **プレプリント ＝ 査読（第三者による内容確認）前の論文原稿**
- **査読前段階で公開することで先取権を主張できることなどから近年広まりつつある**

参考： 科学技術･学術審議会 ジャーナル問題検討部会 第7回（令和2年10月27日） https://www.mext.go.jp/content/20201026-mxt_jyohoka01-000010684_2.pdf

一般的な論文公開までの手続き　巧遅（正確性が高まるが、時間もかかる）



第三者確認なし
誤りがある可能性も

公開までのプロセスに概ね数ヶ月、場合により1年以上かかることもある
査読の結果不採録（公開見送り）になるケースも

プレプリントを用いる手続き　拙速＋巧遅（査読前にプレプリントとして公開し、ジャーナル論文の査読プロセスも平行）

プレプリントサーバー
（プレプリントの公開場所）

一般的な論文公開までの手続きと平行し
査読前の原稿（プレプリント）段階で一般公開

原稿が完成した段階ですぐ公開できる
新型コロナウィルス感染症のような一刻を争うケースではこの速報性が役立つ面も

プレプリント分析の背景

- 迅速性・先発性の観点から、今般プレプリント活用が急速に普及
  - 数学・物理・情報系分野では、1990年代から プレプリントサーバ※1「arXiv」を通じたプレプリント公開が普及
  - 今般のコロナ禍を契機に、医学・生物分野でもプレプリントサーバ※1,2の利用が急増

※1 プレプリントの公開場所を「プレプリントサーバ」という ※2 medRxiv および bioRxiv

- プレプリントを用いた COVID-19/SRAS-CoV2 に関する研究動向の俯瞰
  - プレプリントの性質から、査読論文を対象とした分析と比較して より鮮度の高い研究動向が把握できる可能性
    - 加えて、プレプリントは基本的に無償公開されており、ペイウォールの問題も回避可能
  - 複数分野にまたがるため、題目・概要をAI (自然言語処理)で分析することで分野横断的に研究トピックを抽出
  - 所属情報についても抽出・推定を試み、国別の比較まで試行

→ **既存の論文分析と組み合わせることで、より迅速かつ的確な政策立案や投資判断に寄与**

データ分析上の主な注意点

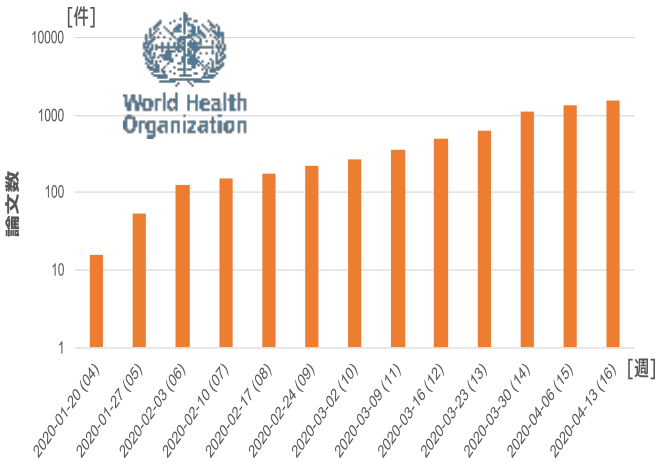
- プレプリントは査読前の論文原稿であり、質に注意が必要
- 研究分野ごとにプレプリントの普及率、投稿先のプレプリントサーバ等が異なる
- 国・地域によってプレプリントの普及率が異なる可能性があり、国際比較に注意
  - 査読論文のような商用データベースなどがなく、プレプリントサーバごとに運営方針や収集・公開情報も異なるため、著者情報（所属等）が捕捉できない場合も多い

論文･プレプリント数の増加傾向（過去の新興感染症との比較）

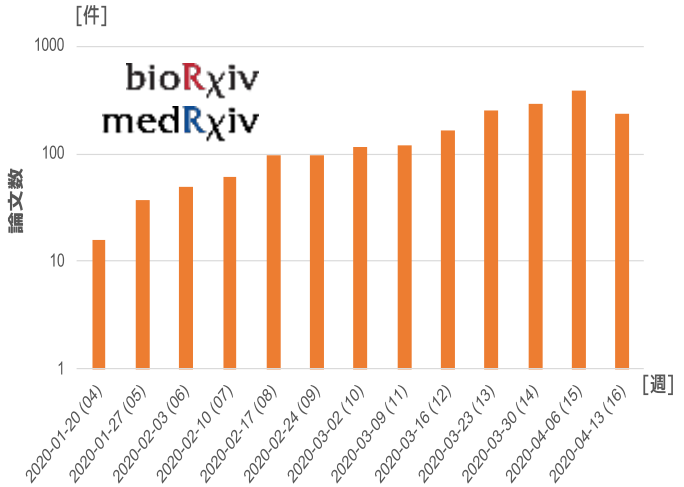
(調査の概要)
調査時点：2020年4月22日時点までに収録された論文データ

調査対象

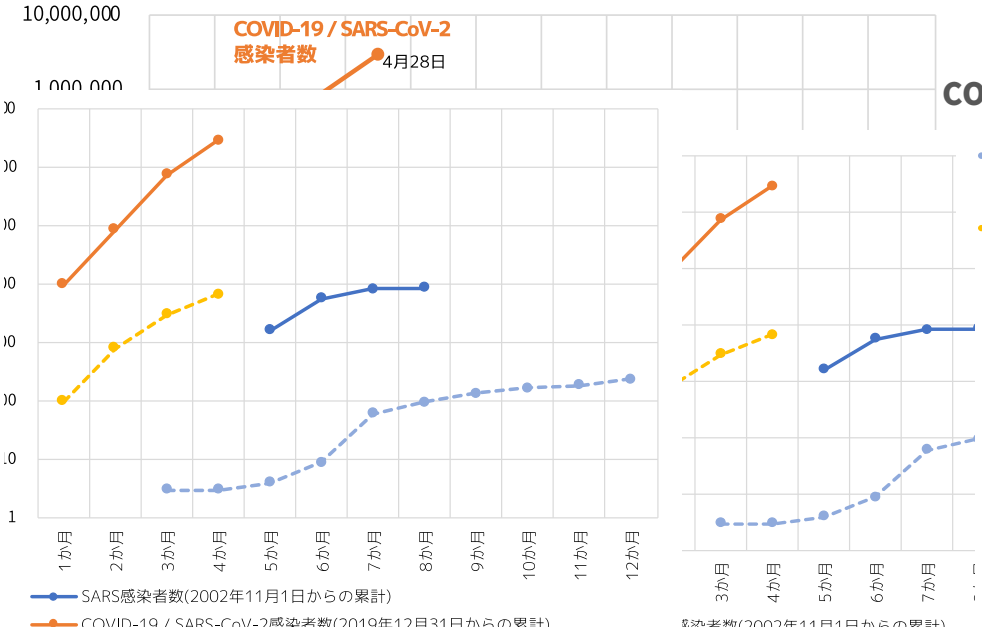
- 世界保健機関(WHO)が公開している文献リスト（ 8,307 件）
- プレプリントサーバ bioRxiv, medRxiv に登録されている文献リスト（ 1,933 件）



世界保健機関(WHO)が公開している文献リスト



プレプリントサーバに登録されている文献リスト



COVID-19 / SARS-CoV-2

感染者数
（2019年12月31日からの累計）

論文数
（DOI登録月 2020年1月からの累計）

SARS

感染者数
（2002年11月1日からの累計）

論文数
（出版月 2002年11月からの累計）

論文･プレプリント数は指数関数的に急増しており、過去の新興感染症 (SARS) と比較してもその増加率は顕著

出典：文科省科政研 (NISTEP) Discussion Paper 181
http://doi.org/10.15108/dp181

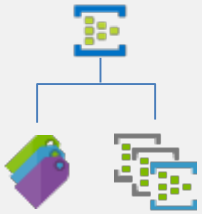
プレプリント分析のワークフロー

データ収集
(Download, Crawling)

APIなどを通じてデータを収集

medRxiv
arXiv.org
bioRxiv
ChemRxiv™
SSRN

データ整理
(タグ付け等)



著者の所属情報などから、国・地域などの情報をタグ付け

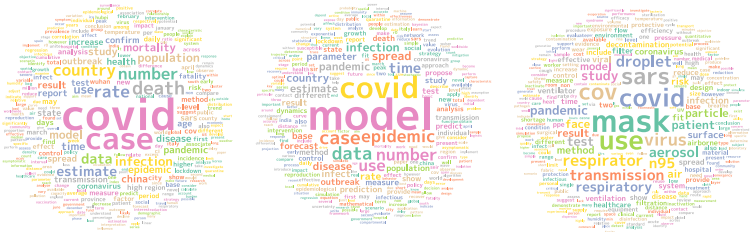
＊ 国・地域は基本的にメールアドレスに基づき整理。
また、基本的に連絡著者1名のみを対象とする。

データ分析

1. **データマイニング (NLP/ML※技術を活用)**
2. **エキスパートジャッジによるトピックの解釈**

※ NLP: Natural Language Processing（自然言語処理）
ML: Machine Learning（機械学習）

データマイニング部分の概要



- **文献の題目・概要から単語を抽出（NLP/ML）**
- **分散表現を算出し，文献を座標値化（NLP/ML）**
- **文献をクラスタリング（NLP/ML）**
- **クラスタごとの頻出語から研究トピックを推定（エキスパートジャッジ）**

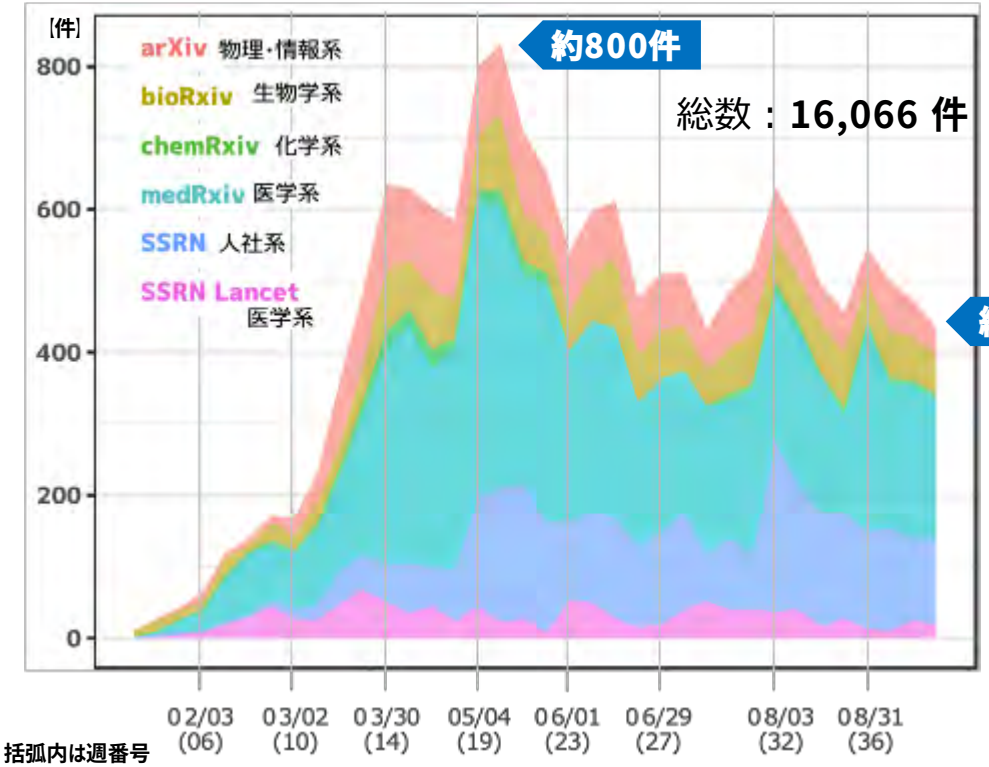
新型コロナウイルス感染症関連の研究動向：
プレプリントを用いた動向把握の試行

- 査読前の論文原稿である“プレプリント”を用い、新型コロナウイルス感染症（COVID-19）関連の研究動向を調査

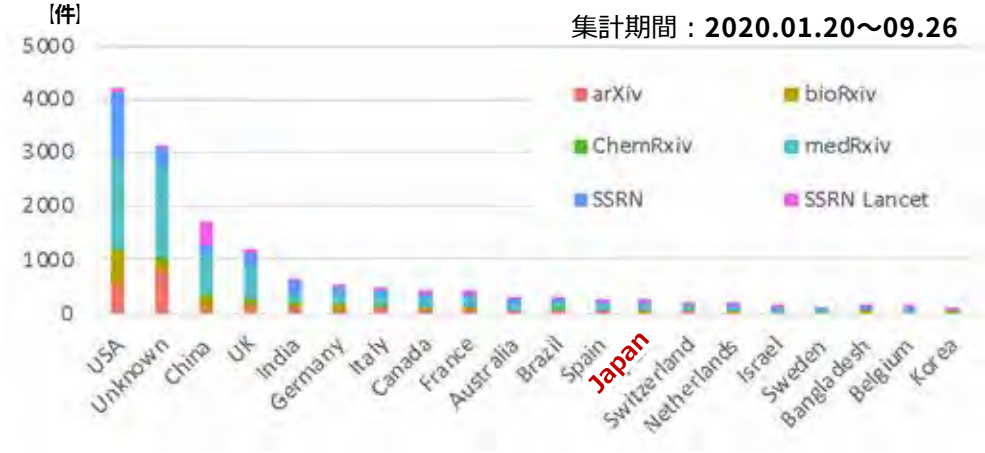
- 投稿件数について
  - 2020年1月20日～9月26日の**約8ヶ月において、16,066 件** の新型コロナウイルス感染症に関連するプレプリントが公開
  - **5月中旬がピーク**で週あたり約800件、7月～9月末は概ね400～500件程度で推移
    - プレプリントは分野ごとに公開場所が異なるが、人社系の割合が徐々に増えている
  - 国・地域別で見ると、**米国の件数が最も多く、**中国、英国、インドと続く
    - 5月頃までは中国が首位であったところ、5月以降は米国が数を伸ばしている

- 研究内容について
  - いわゆるAI技術（自然言語処理）を用い、内容を16分類して専門家が解釈
    - WHOの収集・公開した「ジャーナル論文」を中心としたリストでは見られなかった、“創薬”や“ワクチン開発”に関する話題が出てきている点に特徴
  - 4月頃までは症例報告や国別比較、その後、感染モデル等に話題が推移し、7月以降**公衆衛生や、社会経済に関する内容が増えている**傾向

新型コロナウイルス感染症関連の研究動向（プレプリント分析）：
投稿件数の推移 および 国・地域別の件数比較



括弧内は週番号



総数について

ピークは5月中旬で約800件

9月末には約400件程度に

分野について

医学系の投稿件数が多い

人社系も徐々に増加の傾向
（公衆衛生政策，経済など）

国・地域について※

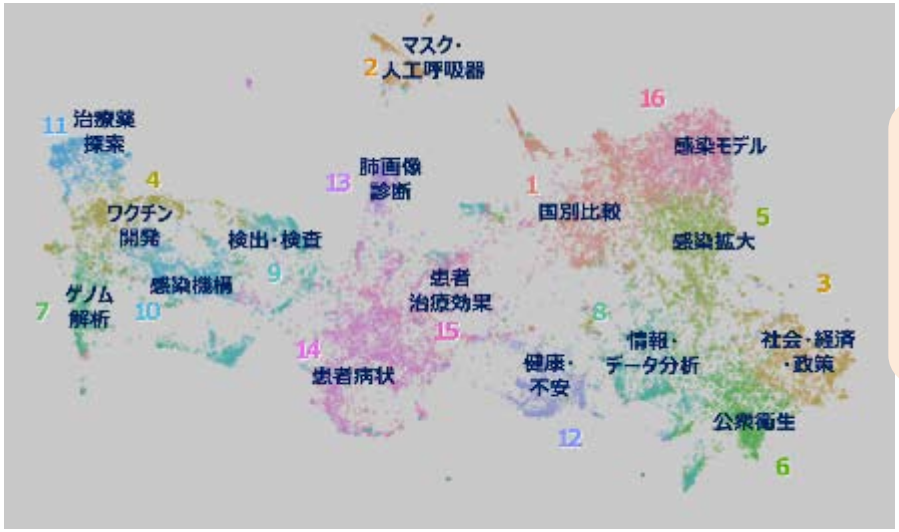
5月までは中国が多かった

以降は米国の数がトップ

※ 連絡著者1名のメールアドレスのみで判定
（一般的な判定方法と異なる）

出典：文科省科政研 (NISTEP) Discussion Paper 186.
http://doi.org/10.15108/dp186

新型コロナウイルス感染症関連の研究動向（プレプリント分析）：
研究の内容に関する分析



出典：文科省科政研 (NISTEP) Discussion Paper 186. http://doi.org/10.15108/dp186

研究内容の分類手法

“いわゆるAI”を用いて分析

記載内容の類似度合いに応じ左図のようにプレプリントを描画

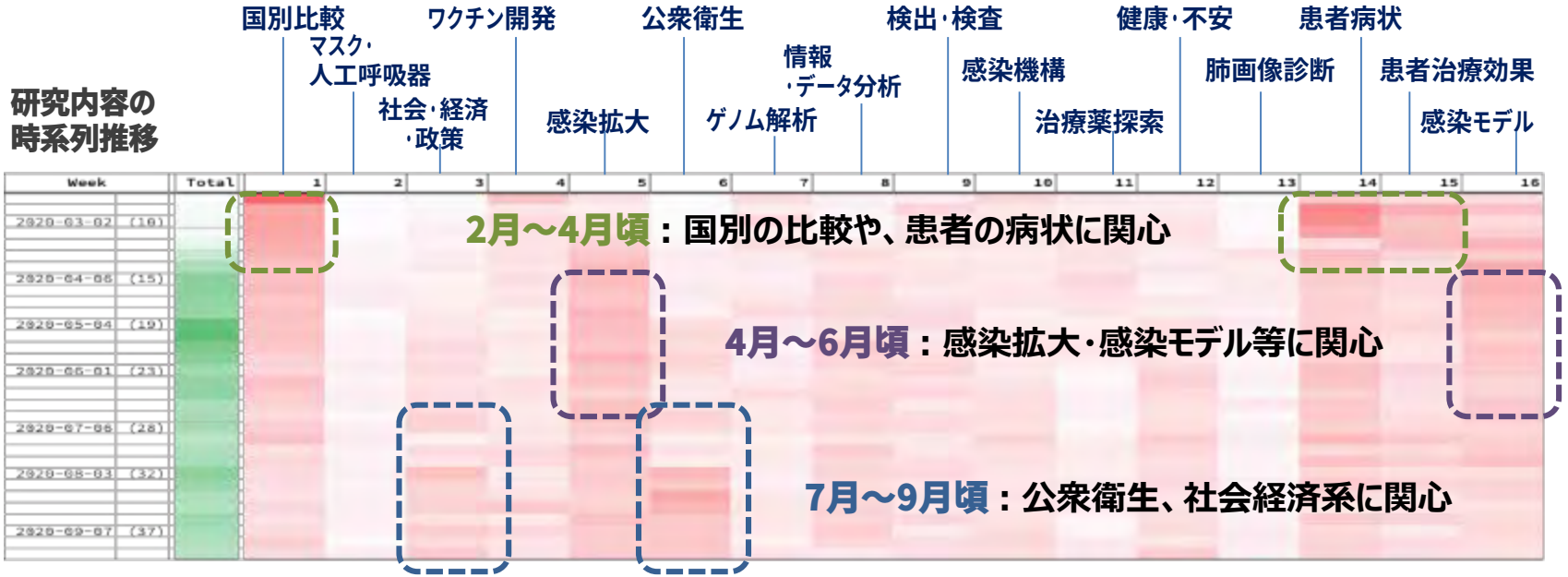
グループ（図中の色が対応）にわけて、専門家が内容を解釈

分類からわかったこと

論文では見えにくかった、“創薬”“ワクチン”に関する話題が見られる

医療と物理・情報系の両方に関わるプレプリントも多く見られている

研究内容の時系列推移



社会経済などの長期的課題へと研究内容のトレンドがシフトしつつある状況がリアルタイムで俯瞰できる

NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY POLICY

『博士人材追跡調査』第３次報告書

本資料は、2020年11月25日に公表した以下の報告書のポイントを示したものです。
「『博士人材追跡調査』第３次報告書」, NISTEP REPORT-188, 文部科学省科学技術・学術政策研究所.
DOI: http://doi.org/10.15108/nr188