DISCUSSION PAPER No.181

COVID-19 / SARS-CoV-2 に関する研究の概況

― 2020 年 4 月時点の論文出版等の国際的なデータからの考察

Summary of research status on COVID-19 / SARS-CoV-2 through an international data around journals and preprints

2020 年 05 月

文部科学省　科学技術・学術政策研究所

小柴　等, 伊神　正貫, 伊藤　裕子,
林　和弘, 重茂　浩美

本DISCUSSION PAPERは、所内での討論に用いるとともに、関係の方々からの御意見を頂くことを目的に作成したものである。

また、本DISCUSSION PAPERの内容は、執筆者の見解に基づいてまとめられたものであり、必ずしも機関の公式の見解を示すものではないことに留意されたい。

The DISCUSSION PAPER series are published for discussion within the National Institute of Science and Technology Policy (NISTEP) as well as receiving comments from the community.

It should be noticed that the opinions in this DISCUSSION PAPER are the sole responsibility of the author(s) and do not necessarily reflect the official views of NISTEP.

【執筆者】

小柴 等	第2調査研究グループ
伊神 正貫	科学技術・学術基盤調査研究室
伊藤 裕子	科学技術予測センター
林 和弘	科学技術予測センター
重茂 浩美	科学技術予測センター

【Authors】

KOSHIBA Hitoshi	2nd Policy-Oriented Research Group, National Institute of Science and Technology Policy (NISTEP), MEXT
IGAMI Masatsura	Research Unit for Science and Technology Analysis and Indicators, National Institute of Science and Technology Policy (NISTEP), MEXT
ITO Yuko	Science and Technology Foresight Center, National Institute of Science and Technology Policy (NISTEP), MEXT
HAYASHI Kazuhiro	Science and Technology Foresight Center, National Institute of Science and Technology Policy (NISTEP), MEXT
OMOE Hiromi	Science and Technology Foresight Center, National Institute of Science and Technology Policy (NISTEP), MEXT

本報告書の引用を行う際には、以下を参考に出典を明記願います。

Please specify reference as the following example when citing this paper.

小柴 等, 伊神 正貫, 伊藤 裕子, 林 和弘, 重茂 浩美「COVID-19 / SARS-CoV-2 に関する研究の概況 ― 2020 年 4 月時点の論文出版等の国際的なデータからの考察」, *NISTEP DISCUSSION PAPER*, No.181, 文部科学省科学技術・学術政策研究所.

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

KOSHIBA Hitoshi, IGAMI Masatsura, ITO Yuko, HAYASHI Kazuhiro, OMOE Hiromi, "Summary of research status on COVID-19 / SARS-CoV-2 through an international data around journals and preprints," *NISTEP DISCUSSION PAPER*, No.181, National Institute of Science and Technology Policy, Tokyo.

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

COVID-19 / SARS-CoV-2 に関する研究の概況
― 2020 年 4 月時点の論文出版等の国際的なデータからの考察

文部科学省　科学技術・学術政策研究所
小柴 等，伊神 正貫，伊藤 裕子，林 和弘，重茂 浩美

要旨

本報では 2020 年 4 月 21 日時点において、世界保健機関（WHO; World Health Organization）が公開している論文データと、プレプリントサーバである bioRxiv, medRxiv でまとめられている論文データを用い、COVID-19 / SARS-CoV-2 に関する研究動向を週単位で調査した。

まず世界における COVID-19 / SARS-CoV-2 の論文数は指数的に伸びており、その伸びは、2002 年の SARS など過去の感染症事例における論文数の増加と比べても特異であることが確認された。現在、世界では COVID-19 / SARS-CoV-2 によってもたらされた危難に対応するために、これまでに例を見ないレベルで研究活動が実施されているといえる。

論文のタイトル・概要に基づくトピック分析から、現在、世界的に研究が実施されていると考えられる 16 のトピック分類を見出した。これらの 16 のトピック分類は、集団発生の確認、積極的な症例の探索などの疫学調査のステップに、よくあてはまることが確認された。また、週単位のトピックの分析から、トピックに表れている時系列的な変化は、疫学調査の段階的な進行状況を反映している可能性が示唆された。これに加えて、国・地域別によるトピックの分布から、感染拡大の時期によって、各国・地域の研究活動の重点が異なる可能性も確認された。

論文数については WHO データにおいて中国と米国が多く、これにイタリア、英国、フランス、ドイツが続いている。日本の論文数は 17 位である。bioRxiv, medRxiv データにおいても中国と米国の論文数が多く、これに英国、イタリア、ドイツ、カナダが続いている。日本の論文数は 8 位である。WHO データにおいて、これら国・地域の論文数と感染者数を分析したところ相関も認められ、感染者数あたりの論文数において、日本は米国、イタリア、英国、フランスよりも高い値を示していることが確認された。

Summary of research status on COVID-19 / SARS-CoV-2 through an international data around journals and preprints

KOSHIBA Hitoshi, IGAMI Masatsura, ITO Yuko, HAYASHI Kazuhiro, OMOE Hiromi
National Institute of Science and Technology Policy (NISTEP), MEXT

ABSTRACT

Since the end of 2019, COVID-19 and the virus SARS-CoV-2 have been spreading globally. And now, countermeasures against these problems have become an urgent issue.

In this report, as of April 21, 2020, we surveyed the publication status of articles by country and region, through the publication data on COVID-19 / SARS-CoV-2 published by the World Health Organization (WHO) and the preprint servers.

First, the number of COVID-19 / SARS-CoV-2 papers in the world is growing exponentially. The growth was also peculiar compared to the increase in the number of articles in past cases of infectious diseases, such as SARS in 2002, which was examined by Scopus. At present, there is an unprecedented level of research activity in the world to respond to the hazards posed by COVID-19 / SARS-CoV-2.

From the topic analysis using word embedding based on the titles and summaries of the papers, we found 16 topic categories that are considered to be currently undergoing research worldwide. These 16 topic categories were found to be well suited to the steps of infectious disease research, such as confirming outbreaks and searching for active cases.

This suggests that the chronological changes represented in the topic may reflect the gradual progression of infectious disease research. In addition, the distribution of topics by country and region confirmed the possibility that the emphasis of research activities may vary depending on the timing of infection spread.

In terms of the number of papers, the WHO data shows that China and the United States have the largest number of papers, followed by Italy, the United Kingdom, France, and Germany. The number of papers in Japan ranks 17th.

In the bioRxiv and medRxiv data, China and the United States also have a high number of papers, followed by the United Kingdom, Italy, Germany, and Canada. Japan ranks 8th in the number of papers.

In the WHO data, the number of articles and the number of infected persons in these countries and regions were analyzed, and a correlation was found. The number of articles per infected person in Japan was higher than that in the United States, Italy, the United Kingdom, and France.

目次

1　序論

2019 年末ごろから感染症である COVID-19，その原因であるウイルス SARS-CoV-2 が世界的に蔓延し，これらへの対策が喫緊の課題となっている[1)]．

本報では世界保健機関 (WHO; World Health Organization) が公開している WHO COVID-19 Database に収録されている論文データと，プレプリントサーバである bioRxiv, medRxiv でまとめられている COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv に収録されている論文データを中心として，COVID-19 / SARS-CoV-2 に関する論文について「論文の内容についての初期的なトピック分析」および「国・地域別の論文出版概況調査」を行った結果を報告する．なお，WHO および bioRxiv, medRxiv のデータについては 2020 年 4 月 21 日までの論文データに基づく分析である．また，これらに加えて査読済み文献の抄録・引用文献サービスである Scopus 等を用い，過去の感染症事例時における論文数の推移と，今回の事例の比較も行う．

本報の主な目的は，現在，その対策が課題となっている COVID-19 / SARS-CoV-2 について，世界中でどのような研究が行われているかを概観することである．当然ながら，COVID-19 / SARS-CoV-2 の論文数は，その国・地域の感染状況とも関連があり，本報の中で示す国・地域別の論文数の動向は，必ずしもその国・地域の研究力を示している訳ではない点には留意が必要である．

[1)] 本レポートは 2020 年 4 月末に執筆しており，その時点での状況を反映した記述となっている．

2　対象・分析手法

後述するデータセットを対象とし，1. 週次（・月次）単位での論文数の推移調査，2. 論文の第 1 著者第 1 所属からの論文と国・地域紐付け，3. タイトル・概要等からの内容分類と紐付け，4. 時系列や内容，国・地域を軸とした論文数調査，を行う．

まず，データセットの概要について述べる．

本報では，WHO が Web サイトで公開している文献・論文データ WHO COVID-19 Database [2] と，プレプリントサーバ bioRxiv および medRxiv [3]が共同で公開している COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv [4] を対象とした．

WHO によると COVID-19 Database は，COVID-19 / SARS-CoV-2 に関する文献の包括的な多言語ソース[5]であり，現時点で入手可能な最も網羅的な情報源である．

WHO のリストには査読を経たジャーナル論文が多く収録されている．そのため，ジャーナル論文については一定の信頼性が担保されているものの，査読の期間を要する分，速報性には劣る．ただし New England Journal of Medicine などのジャーナルで，2–3 日程度の迅速な査読を経て公開に至った論文も含まれる．また，WHO のリストにはジャーナル論文の他，Science 誌に掲載されたコラムのようなものや，シンポジウムの講演概要なども含まれており，全てがジャーナル論文ではない点などに注意が必要である．

bioRxiv，medRxiv はプレプリントサーバであるため，収録されている原稿はあくまで草稿であり，査読を経ておらず信頼性については留意が必要である[6]．他方で，速報性には勝ると考えられ，COVID-19 / SARS-CoV-2 のような緊急性の高いトピックの情報共有には一定のメリットがある．

これら 2 種類の対象について，2020 年 4 月 22 日にデータを収集し，結果として 2020 年 4 月 21 日までの論文データを得た．

最後に，後述する所属データの表記揺れやデータの不完全性から，全著者のデータを網羅することが困難であるため，本報では第 1 著者の第 1 所属のみをカウント対象にする．すなわち，「小柴（NISTEP，はこだて未来大，産総研），伊神（NISTEP）」という著者および所属情報の論文があったとき，小柴の NISTEP のみを対象とし，この場合 NISTEP が日本の組織であるため，日本の論文とカウントする．このため，国際的な協力のもとに執筆された論文については，第 1 著者の第 1 所属の国・地域のみが計数結果に反映される．

また，後述するとおり bioRxiv, medRxiv では，国・地域の推定にメールアドレスを用いる．そ

---

[2] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov (accessed: 2020-04-24)

[3] これらプレプリントサーバの位置づけや動向については文献 [林 20] に詳しい．

[4] https://connect.medrxiv.org/relate/content/181 (accessed: 2020-04-24)

[5] 原文では以下のような表現となっている “This database represents a comprehensive multilingual source of current literature on the topic.”

[6] bioRxiv, medRxiv においても次の注意がなされている．“A reminder: these are preliminary reports that have not been peer-reviewed. They should not be regarded as conclusive, guide clinical practice/health-related behavior, or be reported in news media as established information.”

のため，本報における国地域は，ISO 3166-1 alpha-2 をベースとした国別コードトップレベルドメインを採用する．その結果，たとえば香港は中国には含めず地域として取り扱う[7)]．

2.1 WHO データ

WHO COVID-19 Database に含まれる論文数[8)]は，データ収集時点（2020 年 4 月 22 日）で 8,307 件であり，各論文についてタイトル，概要，著者，DOI などの情報が収録されている．ただし，著者の所属についての情報は含まれておらず，以下に示すように別の情報源から取得する必要がある．

2.1.1 所属データの取得・収集

各論文の国・地域の推定に必要となる著者所属データの取得・収集方法は以下の通りである．

一般的に DOI に関しては Crossref [9)] が提供する Crossref REST API [10)] を用いると，論文タイトル，著者・所属，雑誌名や公開日などの情報を得ることができる．そこでまず，この API を通じて所属データを取得する．ただし，DOI のメタ情報として所属情報等が含まれない場合も多いので，以下に述べる情報源も併用した．

1. 査読済み文献の抄録・引用文献サービスである Scopus において「COVID [11)]」のキーワードで検索し，WHO のリストと付き合わせて所属情報を取得
2. それでも所属が得られなかった論文のうち，Elsevier や Springer など大手出版社の論文について，DOI から各論文にアクセスすることで手作業で所属情報を取得

2.1.2 国・地域の推定

著者所属情報には基本的に国・地域名が記載されており，これを収集することで行う．

ただし，都市名や組織名までしか記載が無かったり，著者の肩書きなど所属ではないデータが記載されているものもある．前者については，著者らが手作業で国・地域名を検索し割り振った．ただし，「Georgia」や「Colombia」のように州や都市などの名前か，国・地域名か判断が難しいものもあり，必ずしも 100% の精度は保証されない．後者については，推定の手がかりがないため除外した．

---

7) 分析対象とする国・地域を合わせるために，WHO データの分析においても香港は中国には含めず地域として取り扱った．

8) 先に述べたように，WHO COVID-19 Database にはジャーナル論文以外も収録されているが以降の議論では一括して論文と記述する．

9) `https://www.crossref.org/`

10) `https://github.com/CrossRef/rest-api-doc` (accessed: 2020-04-24)

11) 他にも適当なキーワードがある可能性があるが，本報においては，速報性を重視し「COVID」を用いた．

2.1.3　有効データ数

上記の手続きに従って処置した結果，WHO COVID-19 Database から取得した論文 8,307 件に対して，国・地域の推定が行えた有効データ数は 4,666 件，約 56% のカバー率となった．

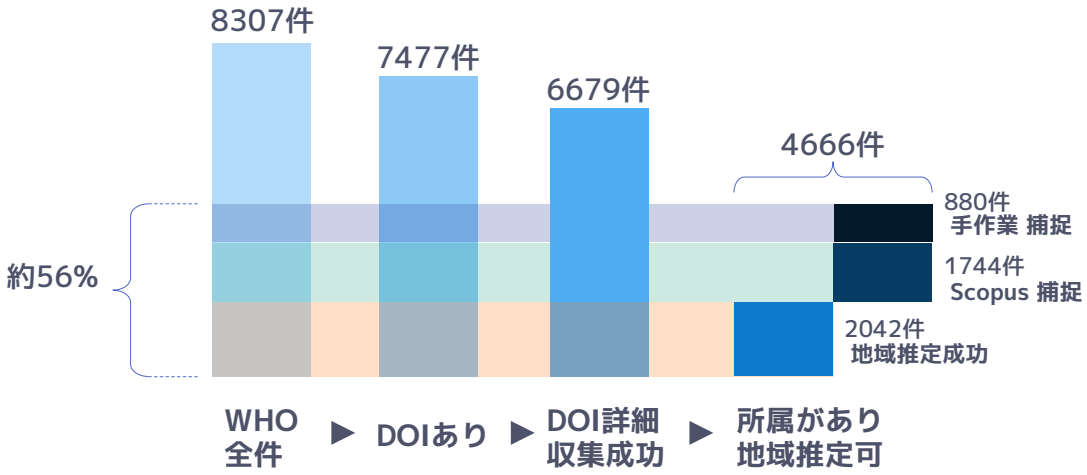


図 1　WHO データの概要

2.2　bioRxiv，medRxiv データ

bioRxiv，medRxiv に含まれる COVID-19 / SARS-CoV-2 に関する論文数は 1,933 件であり，各論文についてタイトル，概要，著者，DOI などの情報が収録されている．ただし，著者の所属についての情報は含まれておらず，以下に示すように別の情報源から取得する必要がある．

2.2.1　所属データの取得・収集

bioRxiv，medRxiv 上の各論文のページにおいて，タイトル，概要，著者・所属，連絡先メールアドレス，本文 PDF へのリンクなどが公開されている．そこでこれらのデータを収集・整理する．

ところで，WHO のデータ収集で述べたとおり，著者所属は自然言語で記入されるため自由度が高く，必ずしも国・地域名が記載されておらず，記載がある場合のパターンも一定しない．したがって，所属データの取得・収集に際して，手作業が必要となり負荷が大きい．

ここで，メールアドレスも著者所属を示す重要なデータである．また，メールアドレスは規約に基づいて設定されるため，機械的に処理がしやすい．そこで，bioRxiv，medRxiv データについては著者所属情報としてメールアドレスを用いることにした．

なお，bioRxiv，medRxiv から取得した論文 1,933 件に対して，連絡先メールアドレスの設定があるものは 1,930 件であった．

2.2.2　国・地域の推定

各論文の国・地域の推定方法は以下の通りである．

bioRxiv, medRxiv については基本的にメールアドレスのトップレベルドメインを用いて各論文の国・地域を推定する．たとえば XXXX@nistep.go.jp のトップレベルドメインは “.jp” で，日本であることが分かる．ただし，“.com” や “.edu”，“.org” など，国とは結びつかないトップレベルドメインも存在する．これらについては，Linux 上の whois コマンドでドメイン登録者の所属国・地域を調べて割り付ける．

なお，gmail.com, yahoo.com, hotmail.com, outlook.com については，ドメイン登録は米国であるものの，利用者が米国在住とは限らない可能性が高いと想定されるため，所属国・地域は不明として取り扱った．

2.2.3　有効データ数

上記の手続きに従って処置した結果，bioRxiv，medRxiv から取得した論文 1,933 件に対して，国・地域の推定が行えた有効データ数は 1,581 件，約 82% のカバー率となった．

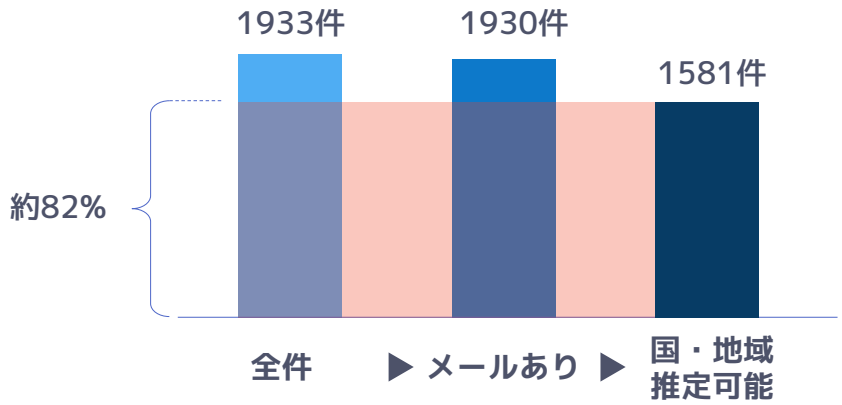


図 2　bioRxiv, medRxiv データの概要

2.3 トピック分析

一口に COVID-19 / SARS-CoV-2 と関連する論文といっても，その内容にはたとえば，公衆衛生や薬学に関するもの，リスクコミュニケーションに関するものなど，さまざまなものが存在すると考えられる．そこで，論文のタイトル，概要に基づいて COVID-19 / SARS-CoV-2 に関係する研究のトピックおよびその変遷や，国・地域毎にその傾向に違いがあるのかについて調べる．

手法としては文献 [小柴 19a, 小柴 19b] を踏襲した．具体的には，論文のタイトル，概要などのテキストに基づいて論文の意味内容を数値データ化・分類した後，それらの関係性を次元圧縮で 2 次元にして可視化するアプローチを取った．これにより，各論文をその意味的な近さに応じて 2 次元空間上に配置・可視化することで，同じトピックの論文の集合を直感的に把握できるようになる．

対象・手続きは以下の通りである．まず，本分析では WHO COVID-19 Database，bioRxiv，medRxiv に含まれる COVID-19 / SARS-CoV-2 に関する全ての論文を母集団とした．その上でタイトルと概要を解析対象として，2 バイトコードを含まないもの，かつ，タイトルと概要を合わせて少なくとも 100 文字以上を有するもの，7,287 件を内容分析の対象とした．

さらに，論文ごとにデータ量の偏りが大きいため，文章中の特徴語を抜き出す手法 (TF-IDF[Sparck72]) を用い，各論文の特徴語上位 20 件までを算出して解析に用いた．分散表現の辞書は，別途 pubMed の 2019 年データ（約 6 千万の論文タイトル・概要データ）を用い，fastText[Bojanowski17, Joulin16] で算出した 300 次元のものを用いた．

3　調査・分析結果

3.1　全体傾向

3.1.1　WHO データ

DOI ベースで発行年月日[12)]が収集できたもの 6,679 件を対象とした論文数の週単位での時系列推移を図 3 に示す．図 3 をみると，片対数グラフに置いて直線的に数が増加していること，つまり世界において COVID-19 / SARS-CoV-2 に関する論文数が，指数関数的に増加していることが確認できる．

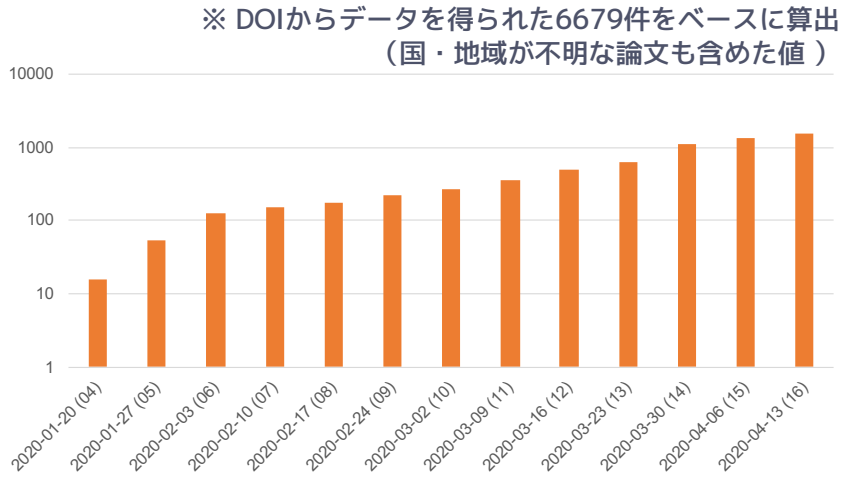


図 3　WHO データにおける論文数の週単位での時系列推移: 括弧内の数字は 1 月第 1 週から数えた週番号

12) 詳細は付録参考のこと

3.1.2　bioRxiv, medRxiv データ

bioRxiv, medRxiv データの全数 1,933 件を対象とした，論文数の週単位での時系列推移を図 4 に示す．図 4 をみると，WHO データと同じく片対数グラフに置いて直線的に数が増加してきていたが，直近の 1 回は論文数が前週を下回っている．この要因については現時点では明確ではなく，今後の推移を見守る必要がある．

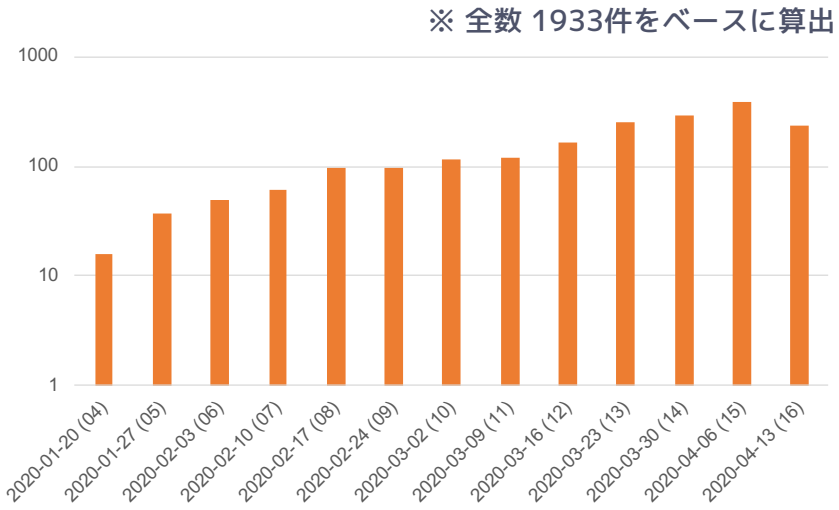


図 4　bioRxiv, medRxiv データにおける論文数の週単位での時系列推移: 括弧内の数字は 1 月第 1 週から数えた週番号

3.2 SARS と COVID-19 / SARS-CoV-2 の論文数等の比較

先に見たように COVID-19 / SARS-CoV-2 の論文数は急増している．この急増傾向は過去の感染症事例と比べて特異なものか調査する．ここでは一例として本報執筆の 18 年前，2002 年に世界的規模で流行したコロナウイルス感染症である SARS（重症急性呼吸器症候群）との比較を示す．

SARS は 2002 年 11 月 16 日の中国の症例から始まった．2003 年 7 月 5 日に WHO によって終息宣言が出されたが，それまでに 32 の国・地域において 8,000 人を超える症例が報告された [感染研 20] [13]．WHO による SARS 感染者数の定期的な報告は 2003 年 3 月 16 日から開始され，2003 年 7 月 5 月まで ほぼ毎日感染者数の報告がなされた [WHO20a]．図 5 において，WHO の報告がはじまった 3 月以降の各月末の SARS 感染者数（累計）を青色の線で示した．SARS 論文数（図 5 の水色の線）は，最初の症例から半年程度は累計でも 10 件程度であったが，2003 年 4 月から 5 月にかけて急上昇した．それでも 2003 年 6 月時点で出版された論文数の累計は 100 件程度である（集計方法については付録 B 参照）．

COVID-19 / SARS-CoV-2 は 2019 年 12 月 31 日の中国の症例に始まり，2020 年 4 月 28 日時点で約 300 万人の症例が報告されている [WHO20b]．WHO による COVID-19 感染者数の定期的な報告は 2020 年 1 月 21 日から開始され，現在も継続中（2020 年 4 月末時点）である．図 5 に 1 月以降の各月末の COVID-19 / SARS-CoV-2 感染者数（累計）をオレンジ色の線で示した．1 月末には 1 万人だった感染者数は，2 月末，3 月末と概ね 10 倍になっており，急激な速度で感染が広がったことが分かる．同じ期間の COVID-19 / SARS-CoV-2 論文数の推移（累計，図 5 の黄色の線）をみると，2020 年 1 月時点でも既に 100 件を超えており，2020 年 4 月 21 日時点で 1 万件のオーダーに迫っている．

このように，SARS と COVID-19 を比較すると，WHO による報告開始のタイミング，感染者数の増加の度合い，論文数の増加の度合いのいずれも，大きな違いがあることが分かる[14]．

現在，世界では COVID-19 / SARS-CoV-2 によってもたらされた危難に対応するために，これまでに例を見ないレベルで研究活動が実施されているといえる．このように活発な研究活動が展開される背景としては，感染者数の規模に加えて，過去の感染症についての知見の蓄積や研究・医療技術の進展に伴う分析等の速度の向上，出版プロセスの電子化に伴う高速化，データの共有による研究の広がりなどの研究活動の高度化・高速化・デジタル化が相互に影響している可能性がある[15]．この点についてはさらなる検証が必要である．

---

[13] 当時，日本人の SARS 感染確認例は報告されていない．

[14] 本報では月次の論文数の変化を示したが，Ebola, Influenza A (H1N1), SARS, MERS, Zika virus の論文数を年次で分析した結果 [Elsevier20] と比較しても，COVID-19 / SARS-CoV-2 の論文数の増加が突出している．

[15] 他方で，SARS から現在までの研究開発の発展にもかかわらず，感染症の脅威は未だ無くならない．この点について，本庶佑京都大学特別教授は「医学は 20 年前に比べても格段に進歩したが新しいウイルスがでてきたら新しい手立てが要る．（中略）たった 1 つの変わったウイルスが出てきて世界がひっくり返るようになる．なんでだと考える人はたくさんいるだろうが，これが現実だ」と指摘している [日経 20].

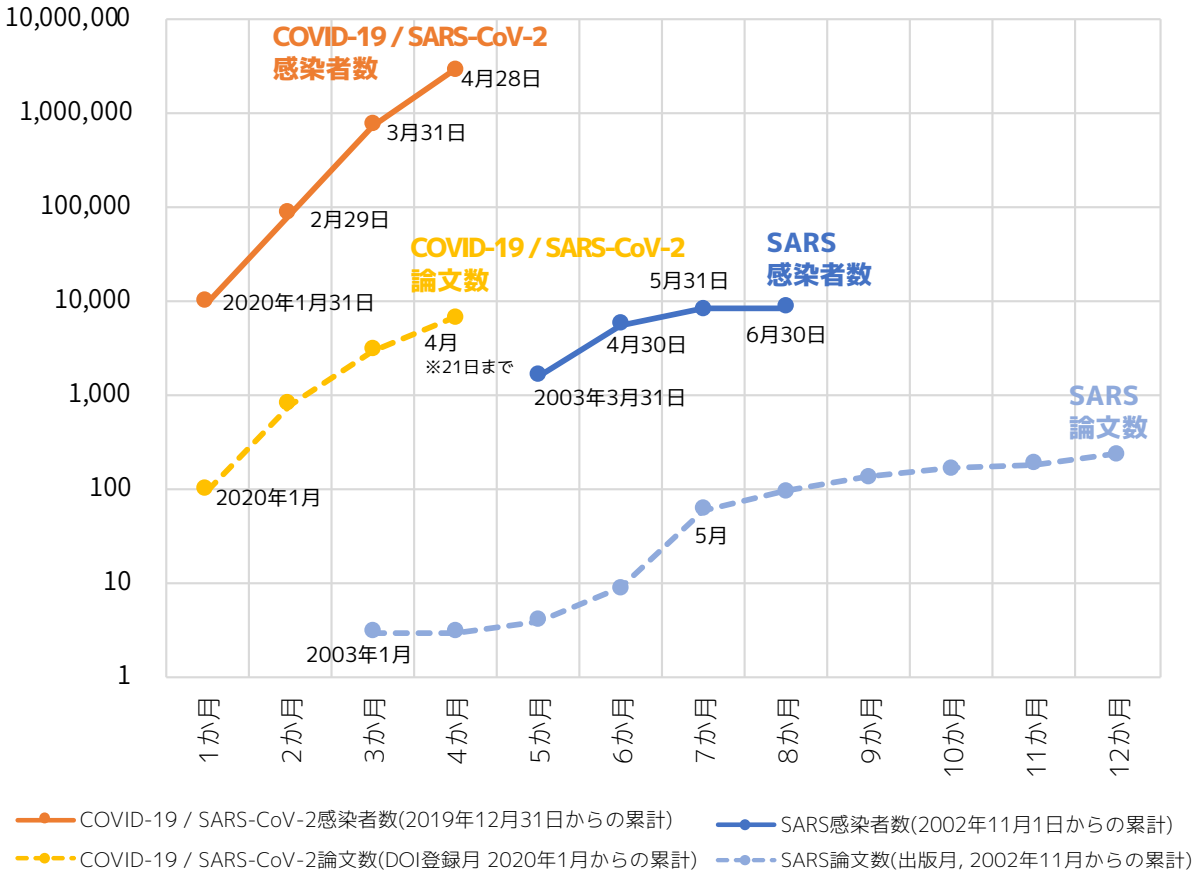


図 5　COVID-19 / SARS-CoV-2 と SARS の感染者数および論文数の比較: 最初の感染が報告された月からの累計，COVID-19 / SARS-CoV-2 の最初の症例は 2019 年 12 月 31 日であるが，2020 年 1 月を 1 か月として集計している．

3.3　論文のタイトル・概要に基づくトピック分析

ここでは WHO データ，bioRxiv，medRxiv データに含まれる COVID-19 / SARS-CoV-2 に関する論文のうち，先に述べた条件（2.3 参照）を満たす 7,287 件について，トピック分析を行った結果を述べる．

分析対象とした論文のタイトルや概要中の単語に対して，レマタイズ[16)]などの下処理を行った上で，分散表現によって意味内容を数値化した後，k-means++ [Arthur07] で 16 のトピックに分類，UMAP[McInnes18] で 2 次元化して可視化した結果を図 6 に示す．

さらに，16 のトピック分類それぞれについて，そこで出現する単語の頻度を用いて作成したワードクラウドを図 7, 8, 9 に示す．ワードクラウドでは出現頻度が多い単語ほど，大きいフォントで表示されている．

16) 過去形や現在進行形などの語尾変化を元に戻す操作

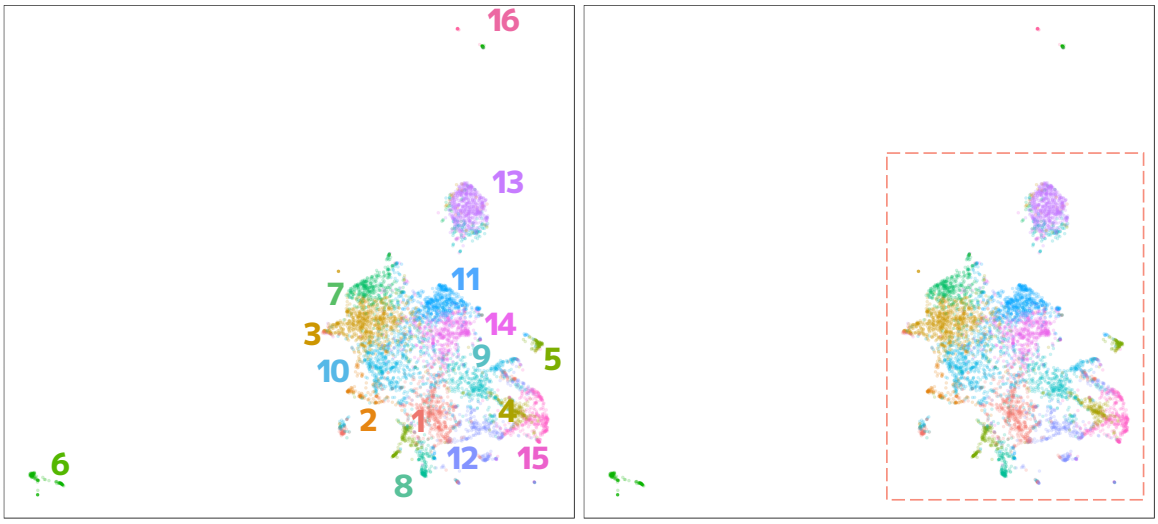


図 6　論文の 300 次元意味空間

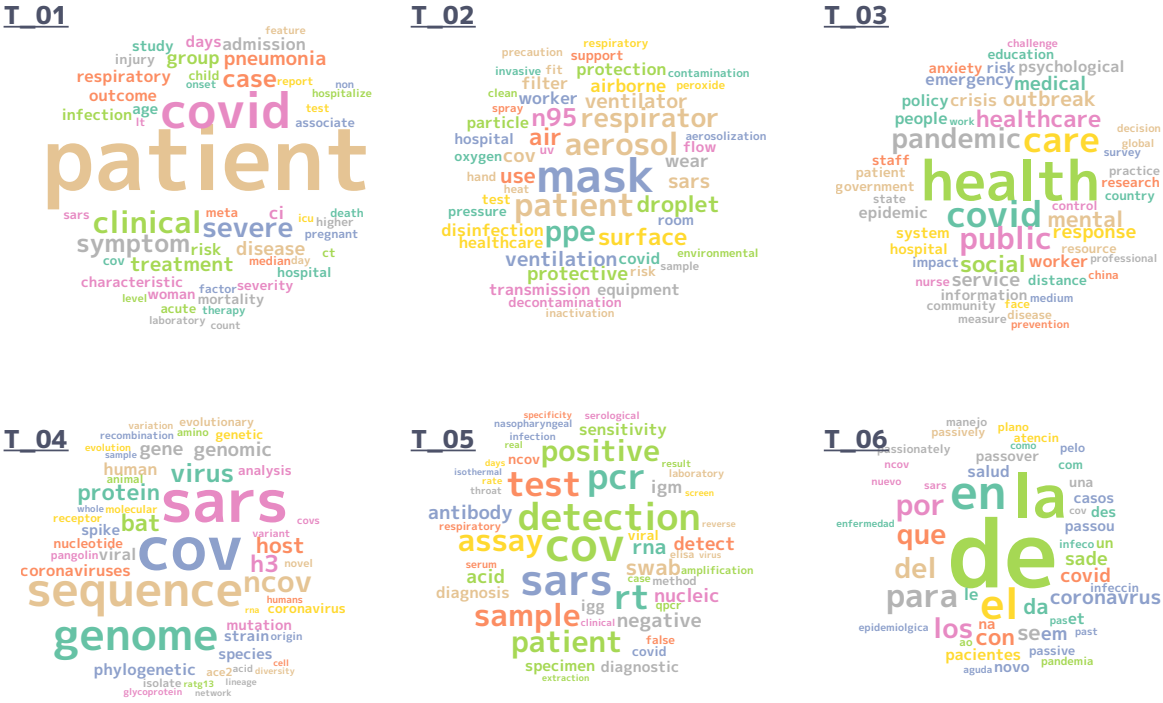


図 7　ワードクラウド (1/3)

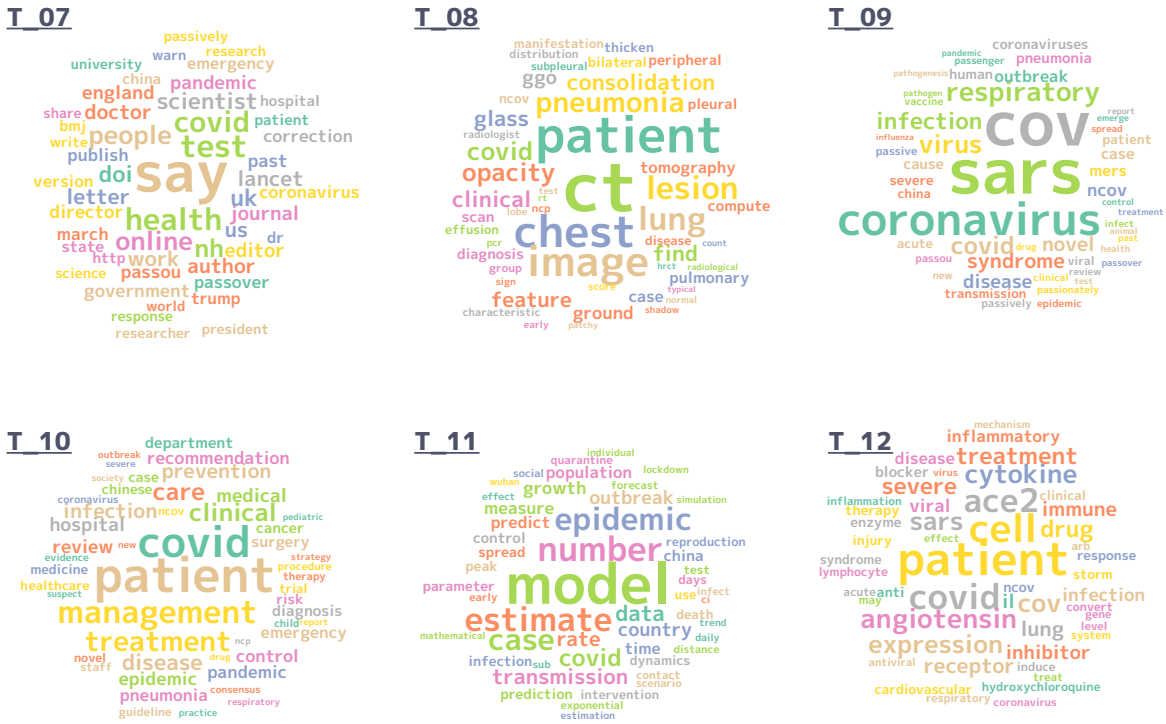


図 8　ワードクラウド (2/3)

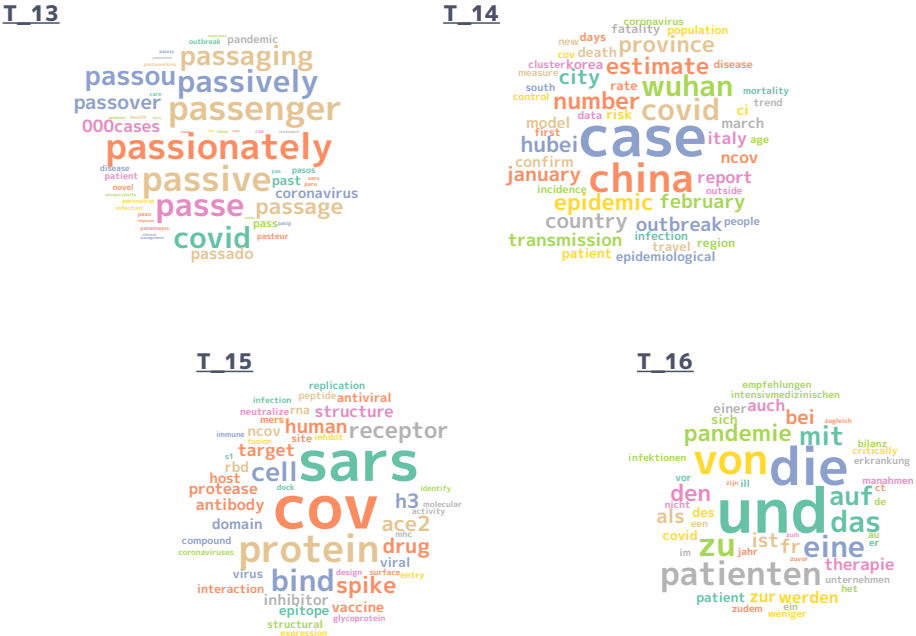


図 9　ワードクラウド (3/3)

各論文の中身まで踏み込まず，単に図 7, 8, 9 で示したワードクラウドの単語からの初期的な解釈を表 1 に示す．表 1 ではワードクラウドの中で，そのトピックの内容を特徴的に示していると思われる単語[17]を著者が抽出した結果を示している．単語が複数あるトピックについては，それらを包含する単語を山括弧〈〉内に示した．これらはあくまでひとつの解釈であって，他の解釈もあり得る．また分類数や分類手法を変えることなどによって，トピックの内容が大きく変化する可能性があることに留意されたい．

表 1　単語ベースの 16 トピック分類の初期的な解釈

ID	解釈
T_01	患者・臨床・重症・治療 〈COVID-19 臨床事例報告〉
T_02	マスク・エアロゾル 〈COVID-19 感染防御研究〉
T_03	健康管理・公衆衛生〈COVID-19 に対する公衆衛生研究〉
T_04	ウイルス・ゲノム解析・シーケンス 〈SARS-CoV-2 ゲノム解析研究〉
T_05	患者検体・PCR 検査・分析・検出 〈SARS-CoV-2 検出法開発研究〉
T_06	（独仏語等の集合）
T_07	リスクコミュニケーション
T_08	患者・胸部 CT 画像・肺炎 〈COVID-19 診断法開発研究〉
T_09	コロナウイルス感染症（SARS・新型コロナ）・呼吸器感染症 〈COVID-19 臨床研究：SARS との比較等〉
T_10	患者・マネジメント・治療・病院 〈COVID-19 の看護研究〉
T_11	感染・数・モデル・推計 〈COVID-19 の感染伝播モデル研究〉
T_12	患者・細胞・ase2・サイトカイン・免疫 〈COVID-19 の病原性発現機構研究〉
T_13	乗客・越境管理
T_14	事例・中国・武漢 〈COVID-19 発見及び臨床事例報告〉
T_15	SARS・新型コロナウイルス・タンパク質・結合・スパイク 〈SARS-CoV-2 感染機構研究〉
T_16	（独仏語等の集合）

また，これら，16 のトピック分類それぞれの時系列変化について図 10 に示す．

分類手法の特性上，各論文は 16 のトピック分類のうち 1 つのみに結びつけられるため，図 10 における各トピックの割合は同一期間内の論文のうち何割がそのトピックに結びついていたかを示す（同一期間の T_01 から T_16 の割合の合計が 100% となる）．また，概要等があっても日時が取得できなかった論文についても併せて計算しているため，参考としてそれらの日時が不明な論文群の集約値についても記載した．図 10 をみると，それぞれの分類の論文が全体に占める割合には，時系列的な変化があることが読み取れる．

[17] ワードクラウドは 1 単語を単位として構成しているので，一部，著者が複数単語の組合せから解釈を示しているトピックもある．

Date	Week	Count			Topic															
		Total	WHO	Rxiv	T_01	T_02	T_03	T_04	T_05	T_06	T_07	T_08	T_09	T_10	T_11	T_12	T_13	T_14	T_15	T_16
2020-01-20	4	17	8	9	0.0%	0.0%	5.9%	23.5%	5.9%	0.0%	5.9%	0.0%	17.6%	0.0%	5.9%	0.0%	5.9%	23.5%	5.9%	0.0%
2020-01-27	5	55	36	19	0.0%	0.0%	7.3%	9.1%	1.8%	1.8%	10.9%	0.0%	12.7%	1.8%	9.1%	1.8%	14.5%	14.5%	14.5%	0.0%
2020-02-03	6	118	89	29	4.2%	1.7%	5.9%	16.1%	1.7%	0.0%	5.1%	3.4%	11.0%	5.9%	6.8%	3.4%	11.0%	16.1%	6.8%	0.8%
2020-02-10	7	131	93	38	4.6%	0.8%	6.9%	9.2%	3.1%	0.0%	6.1%	2.3%	12.2%	3.8%	13.7%	3.1%	3.8%	20.6%	9.9%	0.0%
2020-02-17	8	194	111	83	4.1%	0.5%	6.2%	7.2%	5.7%	0.5%	3.6%	2.6%	16.0%	2.6%	12.9%	5.2%	6.2%	16.0%	10.8%	0.0%
2020-02-24	9	249	156	93	12.9%	0.4%	7.2%	4.0%	6.0%	0.0%	3.6%	4.0%	13.3%	2.4%	10.4%	5.6%	6.8%	17.3%	6.0%	0.0%
2020-03-02	10	281	180	101	8.2%	2.8%	7.1%	5.3%	5.7%	1.1%	7.8%	5.7%	7.8%	5.7%	9.3%	3.6%	7.8%	13.5%	8.2%	0.4%
2020-03-09	11	355	252	103	11.3%	1.7%	8.5%	3.7%	4.2%	0.6%	8.5%	1.7%	11.8%	5.1%	9.0%	3.7%	7.3%	14.9%	7.3%	0.8%
2020-03-16	12	454	321	133	9.9%	1.5%	9.7%	4.2%	3.3%	0.7%	5.5%	2.6%	11.9%	10.4%	7.3%	4.0%	12.1%	10.4%	6.4%	0.2%
2020-03-23	13	539	375	164	13.2%	1.3%	14.1%	2.4%	3.3%	0.7%	6.7%	1.9%	6.1%	8.3%	8.9%	5.0%	13.4%	8.7%	5.0%	0.9%
2020-03-30	14	918	673	245	7.0%	2.2%	16.6%	2.2%	3.7%	3.4%	5.9%	2.1%	6.8%	11.0%	11.2%	5.1%	12.1%	6.6%	3.9%	0.3%
2020-04-06	15	1197	811	386	10.1%	3.0%	15.4%	2.3%	3.6%	1.2%	5.3%	1.7%	4.6%	11.5%	12.2%	5.6%	11.1%	6.5%	5.4%	0.5%
2020-04-13	16	1384	948	436	9.9%	3.8%	15.5%	1.5%	4.0%	1.9%	4.8%	1.7%	5.3%	10.4%	11.7%	6.8%	10.0%	7.6%	4.7%	0.4%
日時不明	不明	1176	1176	0	9.1%	2.6%	9.4%	1.0%	3.0%	10.6%	11.2%	3.1%	8.0%	19.5%	2.9%	5.5%	6.3%	6.0%	1.9%	0.0%

図 10　16 のトピック分類の時系列変化

具体的には，T_04（SARS-CoV-2 ゲノム解析研究），09（COVID-19 臨床研究：SARS との比較等），14（COVID-19 発見及び臨床事例報告） は論文の割合が高まっている時期が早い（2020 年 1 月末〜2 月初め）．また，T_01（COVID-19 臨床事例報告），05（SARS-CoV-2 検出法開発研究），08（COVID-19 診断法開発研究） はそれから少し後の時期（2 月中旬）に割合が高くなっている．さらに，T_03（COVID-19 に対する公衆衛生研究）はそれよりもさらに遅れた時期（3 月下旬）に論文の割合が高くなっている．T_11（COVID-19 の感染伝播モデル研究）には 2 月中旬と 4 月中旬頃の 2 つのピークが出現している．

このように，1 月末から 2 月初めまでに COVID-19 の臨床事例が多数報告され，同時に，感染源を特定するためにウイルスゲノム解析が実施されていたことがわかる．ゲノム解析の結果はウイルスの検出法等の基礎となった．さらに，2 月中旬には早くも SARS-CoV-2 の検出法や診断法に関する論文が多く報告され，COVID-19 の臨床現場（病院）での患者の取扱いや処置に必須かつ重要な知見が集まっていたことがわかる．3 月下旬以降は，個人の治療から集団（社会や各国・地域全体）の治療や健康対策へと感染フェーズが変化したことにより，公衆衛生に関する研究報告が多くなっている．T_11（COVID-19 の感染伝播モデル研究）に示した感染伝搬モデル研究においてはピークが 2 つ生じているが，最初のピークは早期に集まった臨床事例報告（主に武漢）を基にしたモデルであり，遅れて出たピークはその後の米国や欧州などの臨床事例報告に基づいて作られたモデルと推測される．

なお，新たな感染症が発生した際の疫学調査のおおまかなステップとして，

1. 集団発生の確認（たとえば，T_14（COVID-19 発見及び臨床事例報告））
2. 積極的な症例の探索（T_09（COVID-19 臨床研究：SARS との比較等））
3. 観察調査（T_01（COVID-19 臨床事例報告））
4. 症例群の特徴把握（T_01（COVID-19 臨床事例報告），04（SARS-CoV-2 ゲノム解析研究），05（SARS-CoV-2 検出法開発研究），08（COVID-19 診断法開発研究））
5. 感染源・感染経路・リスク因子の仮説設定（T_05（SARS-CoV-2 検出法開発研究），12（COVID-19 の病原性発現機構研究），13（乗客・越境管理），15（SARS-CoV-2 感染機構研究））

6. 仮説の検証（T_03（COVID-19 に対する公衆衛生研究））
7. 感染拡大の防止策の実践・今後の予防策の提案（T_02（COVID-19 感染防御研究），03（COVID-19 に対する公衆衛生研究），07（リスクコミュニケーション），10（COVID-19 の看護研究），11（COVID-19 の感染伝播モデル研究））

があり，適宜必要な感染対策や疫学研究及び臨床研究を実施していくとされている[18]．上記のステップに 16 のトピック分類を並べてみると，よくあてはまることが示された．このことから，トピック自体に表れている時系列的な変化は，感染症分野における疫学調査の段階的な進行状況や関連する研究の内容の変化を反映していると示唆される．

そういう意味では，今回の分析はこれまで人類が遭遇したことがない新しいタイプの感染症・ウイルスである COVID-19 / SARS-CoV-2 研究の早期の状態を捉えたと考えることができる．COVID-19 / SARS-CoV-2 は 2002 年の SARS と同様のコロナウイルスであるにも関わらず，あたかも未知のウイルス及び感染症のように，SARS の時の知見・経験が臨床ケアや治療にほとんど役に立たっていない．我々はインフルエンザウイルスとの長い闘いにおいて，複数の治療薬やワクチンを手に入れているが，コロナウイルスに対する承認された治療薬はまだ 1 つ[19]でワクチンは開発できていない．この違いは，研究の蓄積の差であると考えられる．COVID-19 / SARS-CoV-2 にはまだわかっていないことが多く，今後の研究が待たれる．

最後に，トピック分析で対象とした論文中において出現頻度が多い 5 か国に，日本，台湾，韓国を加えた 8 の国・地域を対象に，それらの国・地域による論文の意味空間中の分布を可視化した結果を図 11 に示す．また，トピック分析で対象とした論文において出現頻度が上位の 20 の国・地域毎に，16 トピック分類ごとの所属割合について示したものを図 12 に示す．

図 11,12 からは，国・地域によって内容の分布に違いがある可能性が示された．特に，中国と米国には違いが認められる．図 11,12 および図 6 について中国と米国を比較すると，中国は T_01（COVID-19 臨床事例報告），05（SARS-CoV-2 検出法開発研究），08（COVID-19 診断法開発研究），14（COVID-19 発見及び臨床事例報告） が多く，米国が T_02（COVID-19 感染防御研究），03（COVID-19 に対する公衆衛生研究），07（リスクコミュニケーション），11（COVID-19 の感染伝播モデル研究）が多いという特徴がある[20]．前述したトピックの特徴を併せて解釈を試みると，中国の論文は集団発生の確認，観察調査，症例群の特徴把握に関する早めのトピック，米国は感染拡大の防止策の実践・今後の予防策の提案に関する遅めのトピックに強い傾向があるといえる．

これは，中国での感染拡大が早期（1 月）に始まって 3 月 1 日にはピークアウトした一方，米国での爆発的な感染拡大は 3 月中旬以降に生じたことに関係があると考えられる．

トピックベースの論文の内容から，中国は武漢の臨床事例報告・SARS との比較・PCR 検査・CT 診断といった COVID-19 に対抗する基本的な部分の研究に貢献したと考えられる．他方，米国は

---

[18] 疫学調査の基本ステップ, 国立感染症研究所 https://www.niid.go.jp/niid/images/idsc/kikikanri/H28/13-7.pdf (accessed: 2020-04-30) および 文献 [柳川 18]

[19] 論文執筆時の 2020 年 5 月 9 日現在.

[20] 図 12 において, 中国と米国の割合の比を求め, 中国の割合が高い上位 4 トピック, 米国の割合が高い上位 4 トピックを示した.

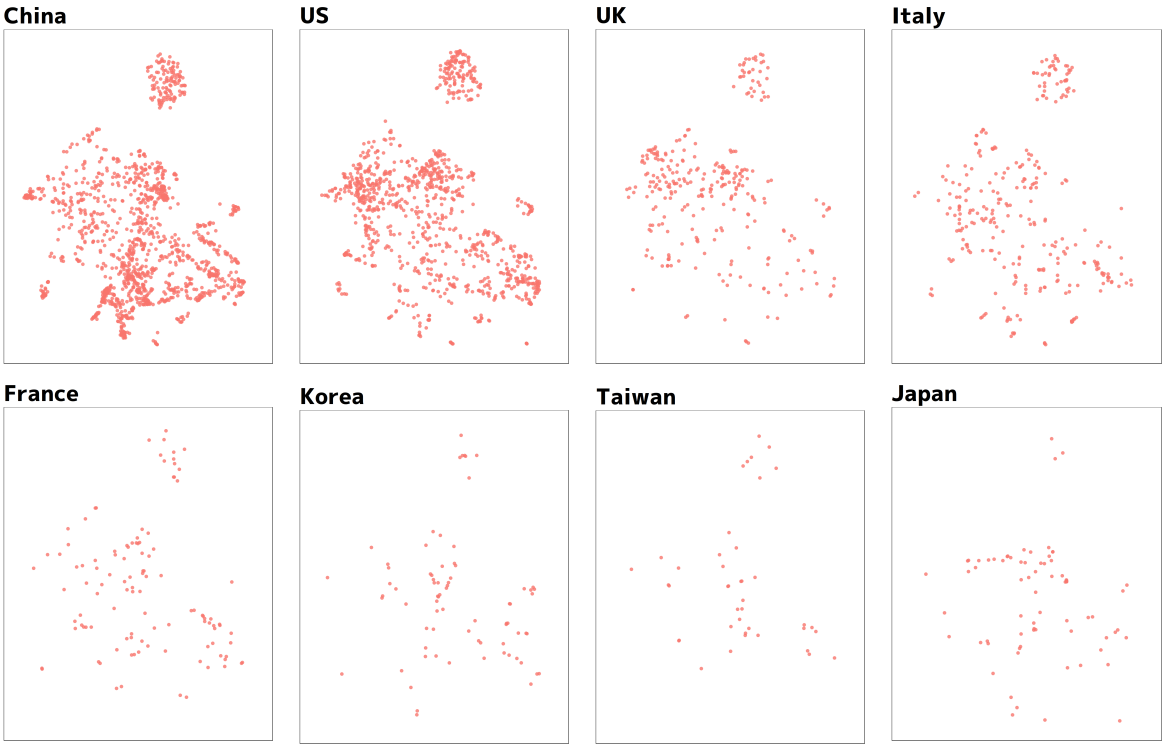


図 11　国・地域別の論文分布

	Name	Count	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_10	T_11	T_12	T_13	T_14	T_15	T_16
1	China	1384	20.3%	1.3%	6.9%	3.8%	7.8%	0.1%	1.5%	4.8%	8.2%	7.5%	6.5%	7.2%	6.5%	11.6%	6.0%	0.1%
2	US	1020	4.8%	3.4%	17.3%	3.9%	2.9%	0.2%	3.4%	1.3%	7.5%	10.8%	12.5%	6.3%	10.9%	7.2%	7.5%	0.0%
3	UK	292	5.1%	4.5%	19.2%	0.7%	2.1%	0.0%	14.0%	0.3%	6.5%	7.5%	16.4%	3.8%	9.2%	8.2%	2.4%	0.0%
4	Italy	286	7.7%	1.4%	9.1%	6.6%	2.4%	0.0%	3.1%	4.9%	5.9%	15.0%	7.3%	9.4%	14.0%	10.5%	2.4%	0.0%
5	France	144	10.4%	0.0%	6.9%	3.5%	2.1%	21.5%	2.1%	1.4%	5.6%	9.0%	10.4%	9.0%	5.6%	7.6%	4.9%	0.0%
6	Germany	128	3.1%	5.5%	10.2%	1.6%	5.5%	0.0%	3.1%	0.0%	7.0%	10.2%	23.4%	1.6%	3.9%	4.7%	7.8%	12.5%
7	Canada	117	6.0%	5.1%	14.5%	6.0%	2.6%	1.7%	1.7%	0.9%	1.7%	17.1%	11.1%	6.0%	12.0%	8.5%	5.1%	0.0%
8	Australia	92	0.0%	5.4%	21.7%	8.7%	3.3%	0.0%	4.3%	0.0%	3.3%	6.5%	9.8%	4.3%	9.8%	18.5%	4.3%	0.0%
9	Hong Kong	91	2.2%	4.4%	15.4%	5.5%	4.4%	0.0%	3.3%	1.1%	7.7%	9.9%	9.9%	1.1%	8.8%	19.8%	6.6%	0.0%
10	India	87	5.7%	2.3%	14.9%	3.4%	1.1%	0.0%	3.4%	1.1%	9.2%	10.3%	14.9%	8.0%	6.9%	5.7%	12.6%	0.0%
11	Spain	77	5.2%	0.0%	9.1%	0.0%	0.0%	39.0%	2.6%	0.0%	3.9%	7.8%	13.0%	5.2%	5.2%	1.3%	7.8%	0.0%
12	Korea	75	8.0%	4.0%	5.3%	5.3%	10.7%	0.0%	0.0%	2.7%	10.7%	5.3%	6.7%	0.0%	9.3%	24.0%	8.0%	0.0%
13	Singapore	73	1.4%	2.7%	8.2%	4.1%	5.5%	1.4%	4.1%	1.4%	15.1%	28.8%	4.1%	0.0%	15.1%	2.7%	5.5%	0.0%
14	Japan	72	8.3%	4.2%	8.3%	1.4%	6.9%	0.0%	2.8%	4.2%	6.9%	2.8%	19.4%	9.7%	4.2%	16.7%	4.2%	0.0%
15	Iran	52	11.5%	3.8%	15.4%	1.9%	0.0%	0.0%	1.9%	1.9%	19.2%	7.7%	7.7%	7.7%	11.5%	7.7%	1.9%	0.0%
16	Switzerland	52	3.8%	3.8%	15.4%	1.9%	1.9%	0.0%	5.8%	3.8%	5.8%	0.0%	23.1%	3.8%	15.4%	9.6%	5.8%	0.0%
17	Brazil	49	2.0%	2.0%	18.4%	2.0%	4.1%	8.2%	0.0%	0.0%	4.1%	4.1%	8.2%	10.2%	18.4%	12.2%	6.1%	0.0%
18	Netherlands	42	2.4%	4.8%	11.9%	0.0%	9.5%	0.0%	4.8%	0.0%	11.9%	4.8%	9.5%	9.5%	11.9%	9.5%	9.5%	0.0%
19	Taiwan	41	7.3%	7.3%	9.8%	4.9%	0.0%	0.0%	0.0%	0.0%	19.5%	0.0%	12.2%	2.4%	14.6%	17.1%	4.9%	0.0%
20	Sweden	30	16.7%	0.0%	6.7%	0.0%	3.3%	0.0%	13.3%	3.3%	13.3%	6.7%	16.7%	6.7%	6.7%	6.7%	0.0%	0.0%

図 12　16 トピック分類ごとの国・地域別論文数

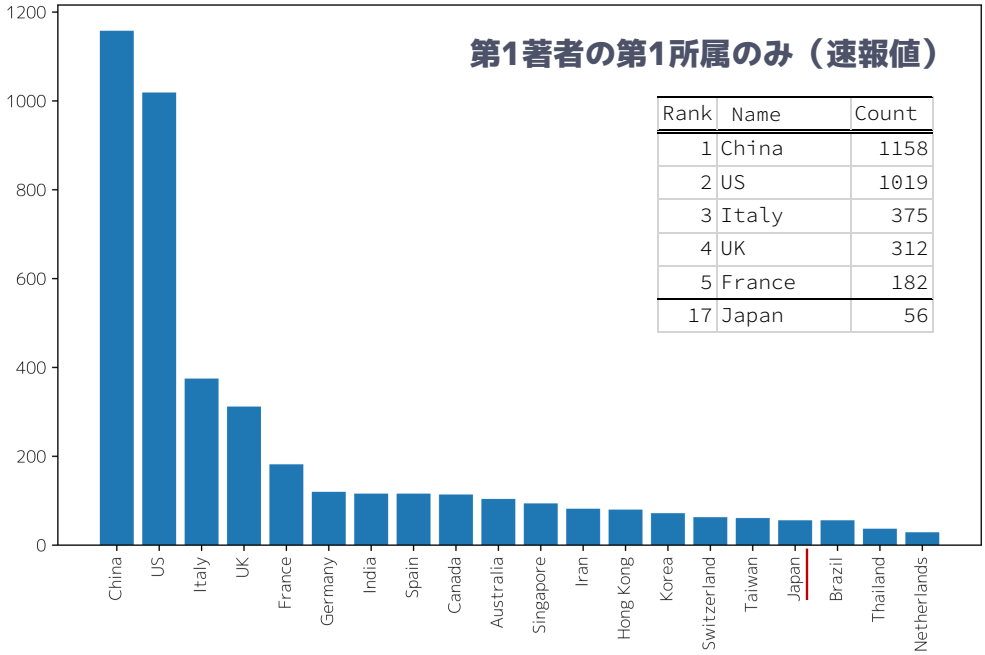
公衆衛生や疫学研究に基づく感染流行モデルの構築など，世界的な終息に向けての研究に貢献していると考えられる．

3.4　国・地域別の論文数

WHO データおよび bioRxiv，medRxiv データのうち，所属機関の情報から国・地域が推定できたものを用いて，国・地域別の論文数を分析した結果を以下に示す．

3.4.1　WHO データ

WHO データにおける，国・地域別の論文数を図 13 に示す．図 13 からは中国と米国が多く，“対数正規分布”や“べき分布”に近い形状となっていることが分かる．中国と米国に続くのは，イタリア，英国，フランス，ドイツとなっており，この中で日本の順位は 17 位である．



Rank	Name	Count
1	China	1158
2	US	1019
3	Italy	375
4	UK	312
5	France	182
17	Japan	56

図 13　WHO データの国・地域別論文数

さらに，国・地域別の期間別の論文数を図 14 に示す．図 14 からは，4 月までは中国の論文数が多数であったところ，4 月以降米国がそれを上回る数となっていることが分かる．

4 月以降に米国の論文数が中国を上回ったのは，感染拡大の状況変化が論文の状況に反映されたためと推測することもできる．たとえば，2020 年 4 月 21 日時点における米国の感染者数は約 79 万件，中国の約 8 万件に比べて約 10 倍である[21)]．一般的に臨床医学の論文数には患者数（感染者数）の多少や推移が影響する傾向がある．

[21)] ここでの感染者数は欧州疾病予防管理センター（ECDC）の公表データに基づく．

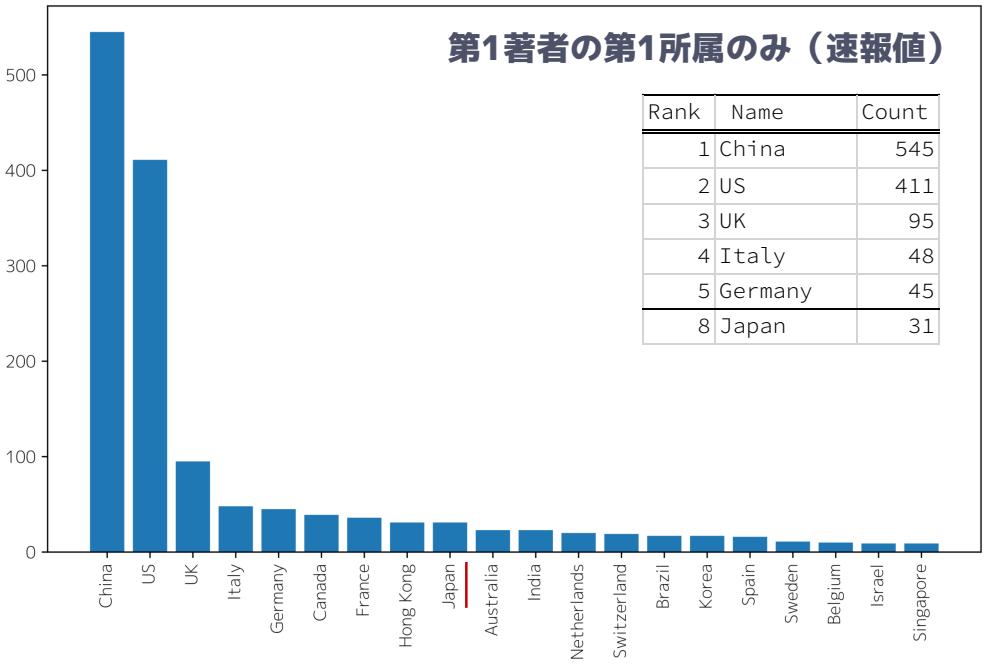
Week	Total	Country/Region															
		China	US	Italy	UK	France	Germany	India	Spain	Canada	Australia	Singapore	Iran	Hong Kong	Korea	Switzerland	Japan
2020-01-20 (04)	16	5	1	0	0	0	1	0	0	0	0	0	0	0	0	0	**0**
2020-01-27 (05)	55	7	6	0	1	1	1	0	0	2	0	0	0	3	0	1	**0**
2020-02-03 (06)	124	30	10	6	5	0	2	2	0	1	1	0	2	0	2	1	**1**
2020-02-10 (07)	151	39	15	1	8	4	3	2	2	3	1	1	0	2	5	1	**1**
2020-02-17 (08)	175	48	20	3	4	0	3	0	3	2	4	1	0	6	5	2	**3**
2020-02-24 (09)	222	60	22	10	15	2	2	2	0	1	1	5	0	3	4	2	**3**
2020-03-02 (10)	271	74	14	7	10	8	3	3	1	3	5	9	1	8	2	7	**3**
2020-03-09 (11)	364	100	49	9	16	10	5	4	8	4	9	5	2	8	5	2	**3**
2020-03-16 (12)	506	130	62	19	23	8	9	5	3	4	12	17	8	7	6	8	**10**
2020-03-23 (13)	641	127	130	42	46	12	15	10	1	19	8	10	4	11	6	6	**3**
2020-03-30 (14)	1138	169	180	75	68	44	25	21	33	22	19	19	17	14	12	7	**9**
2020-04-06 (15)	1330	187	213	88	68	37	21	33	27	19	19	14	23	3	5	6	**9**
2020-04-13 (16)	1525	163	276	108	40	55	28	32	37	30	21	11	22	12	16	18	**10**

図 14　WHO データの国・地域・期間別論文数

3.4.2　bioRxiv，medRxiv データ

次に，bioRxiv，medRxiv データにおける国・地域別の論文数を図 15 に示す．図 15 からは図 13 と同様に中国や米国が多いことが分かる．中国と米国に続くのは，英国，イタリア，ドイツ，カナダとなっており，この中で日本の順位は 8 位（香港と同順位）を占める．

なお，原著論文と違って，プレプリントおよびプレプリントサーバは研究者に広く受け入れられている段階にはなく，掲載論文数も少ないため，この順位の扱いには留意が必要である[22)]．



Rank	Name	Count
1	China	545
2	US	411
3	UK	95
4	Italy	48
5	Germany	45
8	Japan	31

図 15　bioRxiv, medRxiv データの国・地域別論文数

さらに，国・地域別の期間別の論文数を図 16 に示す．図 16 からも，bioRxiv, medRxiv データにおいても WHO データと同様に 4 月までは中国の論文数が多数であったところ，4 月以降米国がそれを上回る数となっていることが分かる．

22) medRxiv の立ち上げは 2019 年 6 月であり，その歴史は浅い

第1著者の第1所属のみ（速報値）

Week	Total	Country/Region														
		China	US	UK	Italy	Germany	Canada	France	Hong Kong	Japan	Australia	India	Netherlands	Switzerland	Brazil	Korea
2020-01-20 (04)	16	8	2	2	1	0	0	0	1	0	0	0	0	1	0	0
2020-01-27 (05)	37	13	8	2	0	0	0	0	2	2	0	0	1	0	0	1
2020-02-03 (06)	49	13	15	2	1	2	1	0	1	2	0	0	0	0	0	0
2020-02-10 (07)	60	25	8	4	0	1	0	1	5	2	1	0	0	1	0	1
2020-02-17 (08)	96	61	10	1	1	1	0	0	1	2	1	0	0	1	0	1
2020-02-24 (09)	96	55	7	4	0	0	1	0	1	2	0	0	0	0	0	2
2020-03-02 (10)	114	58	15	6	1	1	2	1	5	0	1	0	0	2	0	1
2020-03-09 (11)	120	52	19	3	2	2	1	2	3	5	1	2	3	1	0	2
2020-03-16 (12)	162	52	25	9	6	1	3	3	5	4	3	1	2	0	2	4
2020-03-23 (13)	254	77	57	7	5	6	4	3	1	3	4	2	1	4	1	2
2020-03-30 (14)	292	44	74	13	9	14	6	7	3	4	2	5	7	3	6	2
2020-04-06 (15)	389	56	105	20	13	9	15	11	3	3	9	9	2	4	6	0
2020-04-13 (16)	238	30	65	21	9	8	5	8	0	2	1	4	2	2	2	1

図 16　bioRxiv, medRxiv データの国・地域・期間別論文数

3.4.3　感染者数と論文数

ここまでの議論では，単純に論文数によって国・地域の状況を見た．だが，本報で分析対象としている COVID-19 / SARS-CoV-2 については，特に感染者数と論文数との関係が大きいと考えられる．

そこで，ここでは欧州疾病予防管理センター (ECDC) で公開されているデータ[23] に基づく，国・地域別の感染者と論文数の関係について，WHO データにおける論文数上位 30 カ国・地域を対象として図 17 にまとめた[24]．また数値データを表 2 にまとめた．なお表 2 中の “ratio P/C” は論文数を感染者数で割ったものを示している．

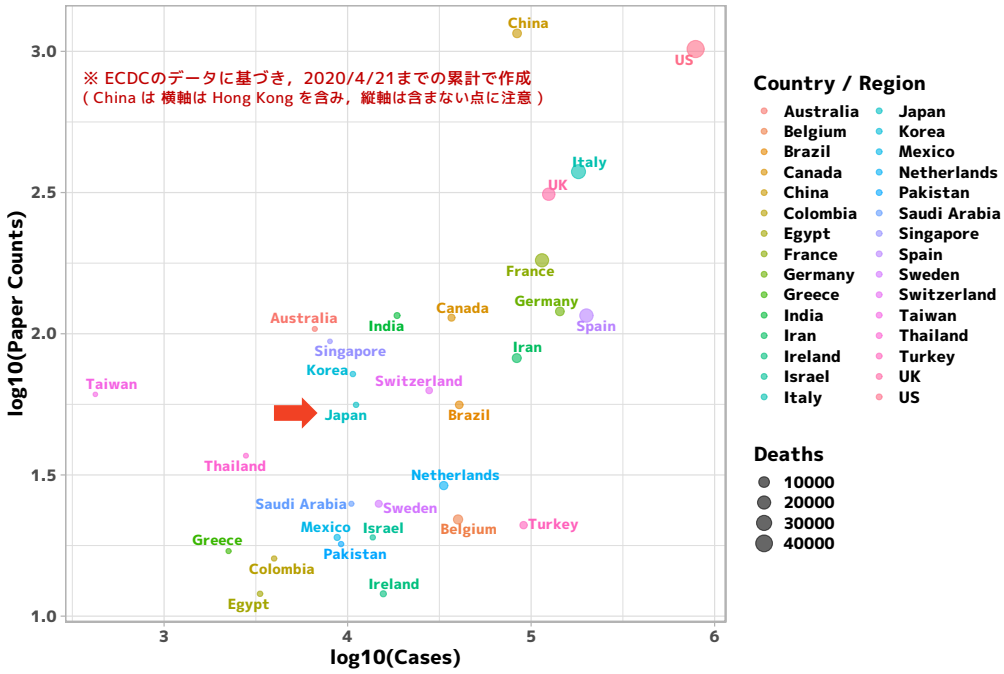


図 17　WHO データの国・地域別論文数と感染者数

図 17 をみると，論文数と感染者数には正の相関がみられ[25]，また，表 2 をみると，感染者数ベースでの論文数において，日本は絶対数 2, 3, 4, 5 位の米国，イタリア，英国，フランスよりも高い比率を示している．

このように論文数については，関係するさまざまな指標で正規化することにより，違った様相が観察できる可能性がある．

---

[23] https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-COVID-19-cases-worldwide (accessed: 2020-04-24)

[24] ECDC では香港を独立に扱っていないため，中国データについてわずかに異なる．

[25] 図に記載の 30 国・地域の範囲で，$r = 0.68$, $p < 0.01$

表 2　WHO データの国・地域別論文数と感染者数データ

Label	Papers	Cases	ratio P/C
China	1158	83849	1.38%
US	1019	787752	0.13%
Italy	375	181228	0.21%
UK	312	124743	0.25%
France	182	114657	0.16%
Germany	120	143457	0.08%
India	116	18600	0.62%
Spain	116	200210	0.06%
Canada	114	36823	0.31%
Australia	104	6625	1.57%
Singapore	94	8014	1.17%
Iran	82	83505	0.10%
Korea	72	10683	0.67%
Switzerland	63	27826	0.23%
Taiwan	61	422	14.45%

Label	Papers	Cases	ratio P/C
Japan	56	11118	0.50%
Brazil	56	40581	0.14%
Thailand	37	2792	1.33%
Netherlands	29	33405	0.09%
Sweden	25	14777	0.17%
Saudi Arabia	25	10484	0.24%
Belgium	22	39983	0.06%
Turkey	21	90980	0.02%
Mexico	19	8772	0.22%
Israel	19	13713	0.14%
Pakistan	18	9216	0.20%
Greece	17	2245	0.76%
Colombia	16	3977	0.40%
Egypt	12	3333	0.36%
Ireland	12	15652	0.08%

4 まとめ

本報では 2020 年 4 月 21 日時点において，世界保健機関 (WHO; World Health Organization) が公開している論文データと，プレプリントサーバである bioRxiv, medRxiv でまとめられている論文データを用い，COVID-19 / SARS-CoV-2 に関する研究動向を週単位で調査した．

まず世界における COVID-19 / SARS-CoV-2 の論文数は指数的に伸びており，その伸びは 2002 年の SARS など過去の感染症事例における論文数の増加と比べても特異であることが確認された．現在，世界では COVID-19 / SARS-CoV-2 によってもたらされた危難に対応するために，これまでに例を見ないレベルで研究活動が実施されているといえる．

論文のタイトル・概要に基づくトピック分析から，現在，世界的に研究が実施されていると考えられる 16 のトピック分類を見出した．これらの 16 のトピック分類は，集団発生の確認，積極的な症例の探索などの疫学調査のステップに，よくあてはまることが確認された．また，週単位のトピックの分析から，トピックに表れている時系列的な変化は疫学調査の段階的な進行状況を反映している可能性が示唆された．これに加えて，国・地域別によるトピックの分布から，感染拡大の時期によって，各国・地域の研究活動の重点が異なる可能性も確認された．

論文数については WHO データにおいて中国と米国が多く，これにイタリア，英国，フランス，ドイツが続いている．日本の論文数は 17 位である．bioRxiv, medRxiv データにおいても中国と米国の論文数が多く，これに英国，イタリア，ドイツ，カナダが続いている．日本の論文数は 8 位である．WHO データにおいて，これら国・地域の論文数と感染者数の関係を調査したところ相関も認められ，感染者数あたりの論文数において，日本は米国，イタリア，英国，フランスよりも高い値を示していることが確認された．

4.1 留意事項

本分析はあくまで簡易的・速報的なものであり，**データの読み取り，活用にはさまざまな注意が必要である．**

まず，著者所属の判定は一部著者らの手作業による判定が含まれる．さらにカウント対象は第 1 著者の第 1 所属のみである．WHO のデータについては国・地域の判定が行えたものは半数程度である．bioRxiv, medRxiv はメールアドレスベースで判定したが，たとえば中国のプロバイダのメールアドレスを使っている米国機関所属の著者がいた場合，中国の著者と誤計上されている可能性もある[26)]．これに加えて，COVID-19 / SARS-CoV-2 の論文数は，その国・地域の感染状況とも関連がある（3.4.3 参照）．

以上のような理由から，本報の中で示した国・地域別の論文数の動向は，必ずしもその国・地域

[26)] WHO データについては，総数約 8.3 千件のうち，国・地域が推定できたものは 4.7 千件とカバー率が 56%，bio/medRxiv データについては総数約 2 千件のうち，国・地域が推定できたものは 1.6 千件，カバー率 82% となった．

の研究力を示している訳ではない点には留意が必要である．

2002 年の SARS との比較についても，感染者数はもちろん，インターネットなど情報通信技術の発展普及状況など，多くの相違点が存在する．さらには SARS 禍における経験の蓄積が活かされた結果として生み出された相違点の存在も考えられる．したがって，それらの考慮なしに単純に数を比較することには注意が必要である．

内容分析に関しては，国・地域によってトピック分布が異なることは示せているが，その意味するところの解釈については慎重な検討を要する．

最後に，すでに示したとおり論文数は対数グラフで直線的に増加している．そのため，本報で述べた状況はあくまで 2020 年 4 月 21 日時点のものであり，今後大きく様相が変わっている可能性もある．

付録 A　WHO データの発行年月日

DOI を通じ，Crossref から取得可能な主な日付データとしては，‘published-online’, ‘published-print’, ‘published’, などの pub 系や ‘created’ などを挙げることができる．

Crossref から取得されたデータには，pub 系の日付データは少なくともひとつが付与されているが，場合により発行年月のみで日が付与されていないものがある．今回の事象は変化が早く，それに対応して週単位での集計が念頭にあるため，分析粒度が月単位では粗すぎる．

‘created’ は DOI の発行年月日であり[27)]，日の情報が基本的に付与されている．その日付は，多くの場合 pub 系の最も古い日付と一致し，場合により pub 系よりも更に古い日付が設定されることもある．

今回は前述した「週単位での集計」のため，WHO データの発行年月日として，「created」を採用した．すでに述べたとおり，‘created’ は公開年月日ではなく，DOI の発行年月日である点には注意が必要である．

なお，bioRxiv, medRxiv については，システムへの投稿年月日 (Posted) データが付与されているため，単にこれを用いる．

[27)] `https://github.com/CrossRef/rest-api-doc/blob/master/api_format.md` (accessed: 2020-04-24)

付録 B　SARS の論文データの取得および集計

本報の SARS の論文データの取得および集計は以下の手順で実施した.

1. Scopus で以下のクエリで論文を検索し，書誌情報をダウンロードした (3,559 件)．検索は 2020 年 4 月 28 日に行った.

```
TITLE-ABS-KEY((SARS AND coronavirus*) ) AND ( LIMIT-TO (
PUBYEAR,2008) OR LIMIT-TO ( PUBYEAR,2007) OR LIMIT-TO (
PUBYEAR,2006) OR LIMIT-TO ( PUBYEAR,2005) OR LIMIT-TO (
PUBYEAR,2004) OR LIMIT-TO ( PUBYEAR,2003) OR LIMIT-TO (
PUBYEAR,2002) )
```

2. 上記でダウンロードした論文のうち DOI を持つものについて Crossref 経由で ‘published-print’，‘created’ の情報等を取得した．Crossref から何らかの情報がダウンロードできた論文の件数は 2,540 件であった.
3. 上記で情報が取得できた論文について，論文の出版年月日の情報である ‘published-print’ を用いて月別で論文数の集計を行った．DOI を後から遡って付与したと思われる論文が見られたため，ここでは DOI の発行年月日である ‘created’ は用いなかった.

参考文献

[Arthur07] Arthur, David and Vassilvitskii, Sergei：K-means++: The Advantages of Careful Seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* pp,1027–1035, 2007. http://dl.acm.org/citation.cfm?id=1283383.1283494

[Bojanowski17] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T.: Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017. arXiv:1607.04606

[Elsevier20] The Elsevier Community： Infographic: global research trends in infectious disease. https://www.elsevier.com/connect/infographic-global-research-trends-in-infectious-disease (accessed 2020-04-29)

[Joulin16] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T.: FastText.zip: Compressing text classification models, *arXiv preprint*, 2016. arXiv:1612.03651

[McInnes18] Leland McInnes, John Healy and James Melville：UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint,* 2018. arXiv:1802.03426

[Sparck72] Sparck Jones, K.： A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation,,* Vol. 28 No. 1, pp. 11-21. 1972. https://doi.org/10.1108/eb026526

[WHO20a] World Health Organization：Severe acute respiratory syndrome (SARS). https://www.who.int/csr/don/archive/disease/severe_acute_respiratory_syndrome/en/ (accessed 2020-04-29)

[WHO20b] World Health Organization： Coronavirus disease (COVID-2019) situation reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed 2020-04-29)

[感染研 20] 国立感染症研究所：SARS（重症急性呼吸器症候群）とは. https://www.niid.go.jp/niid/ja/kansennohanashi/414-sars-intro.html (参照 2020-04-29)

[小柴 19a] 小柴 等, 森川 想：議事録を用いた我が国における議会・行政の関係性分析手法. **人工知能** Vol.34, No.5, pp.E-J47_1-10, Sep 2019. https://doi.org/10.1527/tjsai.E-J47

[小柴 19b] 小柴 等, 池内 健太, 元橋 一之： 日米の特許データと論文データを用いた Mapping Patents の試行. **人工知能学会「社会における *AI* 研究会」** Vol.35, No.8, pp.1–8, Nov 2019. http://id.nii.ac.jp/1004/00010441/

[日経 20] 日本経済新聞：脅威は続く、科学は途上 京都大学特別教授・本庶佑氏 コロナと世界（2）. 2020 年 4 月 10 日朝刊 1 面

[林 20] 林 和弘： MedRxiv, ChemRxiv にみるプレプリントファーストへの変化の兆しと オープンサイエンス時代の研究論文. *STI Horizon 2020* **春号** Vol.6, No.1, Mar 2020. https://doi.org/10.15108/stih.00205

[柳川 18] 柳川 洋：臨床研究と疫学. **月刊地域医学** Vol.32, No.9, pp.804(54) – 812(64), 2018.

DISCUSSION PAPER No.181

COVID-19 / SARS-CoV-2 に関する研究の概況
― 2020 年 4 月時点の論文出版等の国際的なデータからの考察

2020 年 05 月

文部科学省 科学技術・学術政策研究所
小柴 等, 伊神 正貫, 伊藤 裕子, 林 和弘, 重茂 浩美

〒100-0013 東京都千代田区霞が関 3-2-2 中央合同庁舎第 7 号館 東館 16 階
TEL: 03-3581-2391 FAX: 03-3503-3996

Summary of research status on COVID-19 / SARS-CoV-2
through an international data around journals and preprints

May 2020

KOSHIBA Hitoshi, IGAMI Masatsura, ITO Yuko, HAYASHI Kazuhiro, OMOE Hiromi
National Institute of Science and Technology Policy (NISTEP)
Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan

http://doi.org/10.15108/dp181

https://www.nistep.go.jp