

特別企画 第404回 OUGライフサイエンス分科会

～第20回情報プロフェッショナルシンポジウムINFOPRO2023ver.～

医学分野のプレプリントの文献データベース収録

2023年7月6日（木）13:20～13:45

INFOSTA OUG ライフサイエンス分科会

日本オンライン情報検索ユーザー会 Online Users Group(OUG)

OUG ライフサイエンス分科会

<https://www.infosta.or.jp/research/oug-life/>



- 医薬分野から広くライフサイエンス全般を対象に検索演習、勉強会、見学会等を開催
- 開催日：毎月第3木曜午後、年間9回程度開催
- 開催場所：現在はWeb（リアルの場合は日本図書館協会会議室（茅場町））
- 本会はメンバー全員によるグループ（輪番）制で会の企画・運営を行っております。
- OUG入会手続きは入会申込（OUG/SIG）のページをご覧ください。
OUG会員の方で初めてライフサイエンス分科会に出席される場合は、予め事務局経由にて主査までご連絡ください。
- OUG非会員はビジター参加も可能です。

背景・目的

■背景

新型コロナウイルス感染症の拡大とともに医学分野でもプレプリントでの論文公開が増加し、医学学術文献データベースへの収録も開始された

- 2020年6月よりMedline (PubMed) へCovid-19関連を収録開始
- 2021年12月よりEmbasePreprintsが公開
- 2022年3月よりJ-GLOBALで収録開始
- 2022年7月よりJDreamで収録開始
- 2023年1月にMedline (PubMed) へCovid-19以外のプレプリント収録を表明

そこで、プレプリントを対象に文献データベースを検索し、その検索結果から収録状況を調査することとした。

■目的

文献データベースへのプレプリント収録状況を調査・検証し今後の文献データベースの活用に繋げる

プレプリントについて

学術雑誌に投稿し査読を受ける前の論文原稿で査読前論文とも呼ばれる。プレプリントを掲載して公開するプラットフォームがプレプリントサーバー。学術論文の増加と同様にプレプリントも増加し続けている。

【主なプレプリントサーバーの例】

- arXiv：物理学分野。1991年～（途中、LANL preprint archiveから名称変更）
- SSRN：社会学分野。1994年～
- bioRxiv：生物学分野。2013年～
- Research Square：SpringerNatureの出資、2018年～
- medRxiv：医学分野。2019年～

調査内容

■対象データベース

JDreamPREPRINT、J-GLOBAL、EmbasePreprints、MEDLINE

■検索内容：プレプリントを対象に検索

- ① タイトルに「Covid-19」を含み、かつ、出版年2023年
- ② タイトルに「Hypertension」を含み、かつ、出版年2023年

(検索実施日) 2023年3月20日※

※JDreamPREPRINTとJ-GLOBALに更新が反映される日を選定

■調査内容

収録データを比較する：重複状況や各データの中身
相違点があればその要因、また、その後（5月・6月）の収録状況を確認する

調査対象の文献データベースのプレプリントの収録

	JDreamPREPRINT	J-GLOBAL	MEDLINE(PubMed) ※2	EmbasePreprints
作成機関	国立研究開発法人科学技術振興機構 (JST)		アメリカ 国立医学図書館	エルゼビア社
プレプリント収録	2022年7月～	2022年3月～	2020年6月～	2021年12月～
プレプリント収録年	arXiv : 2016年～ bioRxiv : 2019年～ medRxiv : 2019年～		2020年6月～	2021年～ 過去分の遡及あり(6月末時点で1,200件程度)
収録源と収録数※1	約116万件 arXiv:約95万、 bioRxiv:約17万、medRxiv:約4万		約1万件 arXiv: 365、bioRxiv : 5,746、 medRxiv : 2,562、 ResearchSquare : 1,288、 ChemRxiv : 24、SSRN : 27	約7.5万件 bioRxiv: 約58,000 medRxiv: 約17,000
アクセス	JDream (有料)	Webで無料公開	MedlineはDialog、STN、OVIDなどで提供 (有料) PubMedはWebで無料公開	Dialog、STN、OVIDなどで提供 (有料)

※ 1 数字は2023年6月末時点

※ 2 NIHが研究助成したプレプリントが対象で、2020年当初はCovid-19関連、2023年からテーマを限定せず収録。

① 検索式 COVID-19

データベース	検索式
JDream PREPRINT	(COVID-19/TI + COVID-19/TIEN) *(2023-2023/PY)
J-GLOBAL※	タイトル「COVID-19」 発行年「2023」 文献種別「プレプリント」
Embase Preprints (Dialogを使用)	(ti(covid-19)) and (pd(20230101-20231231))
MEDLINE (Dialogを使用)	(ti(covid-19)) and (rtype.exact("Preprint")) and (pd(20230101-20231231))

※J-GLOBALの検索画面

▼ 選択中フィルタ

- 文献種別 : プレプリント ×
- 発行年 : 2023 ×

詳細検索 : 文献 ▼ を探す ? リセット 閉じる 検索

詳細検索を行う基本情報を選んでください

すべて 通常文献 **プレプリント**

フリーワード × 同義語を見る (0)

and ▼ タイトル ▼ COVID-19 × + フィールド内で語句を追加 (ORのみ)

②検索式 hypertension

データベース	検索式
JDream PREPRINT	(hypertension/TI+hypertension/TIEN)*(2023/PY)
J-GLOBAL※	タイトル「hypertension」 発行年「2023」 文献種別「プレプリント」
Embase Preprints (Dialogを使用)	(ti(hypertension)) and (pd(20230101-20231231))
MEDLINE (Dialogを使用)	(ti(hypertension)) and (rtype.exact("Preprint") and pd(20220101-20221231))

※J-GLOBALの検索画面

選択中フィルタ

- 文献種別 : プレプリント ×
- 発行年 : 2023 ×

詳細検索 : 文献 を探す ? リセット 閉じる 検索

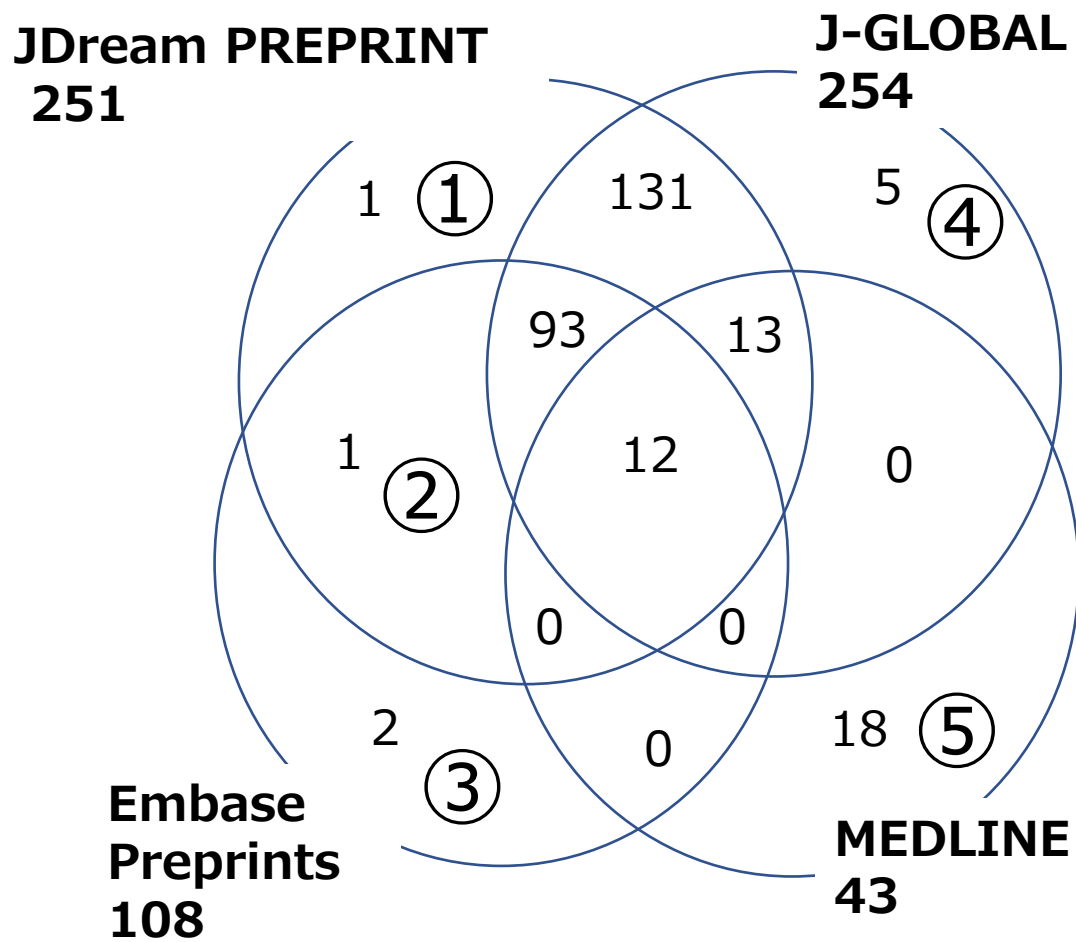
詳細検索を行う基本情報を選んでください

すべて 通常文献 **プレプリント**

フリーワード × 同義語を見る (0)

and ▼ タイトル ▼ hypertension □ × + フィールド内で語句を追加 (ORのみ)

COVID-19 検索結果



データベース	件数
JDream PREPRINT	251
J-GLOBAL	254
Embase Preprints	108
MEDLINE	43

- ① JDreamPREPRINTのみの1件 : arXiv
- ② EmbasePreprintsとJDreamの1件 : medRxiv
- ③ EmbasePreprintsのみの2件 : medRxiv
- ④ J-GLOBALのみの5件 : medRxiv
- ⑤ MEDLINEのみの18件 : medRxiv(6)、
bioRxiv(2)、arXiv(2)、ResearchSquare(8)

- ① JDreamPREPRINTのみの1件 : arXiv
- ② EmbasePreprintsとJDreamPREPRINTの1件 : medRxiv

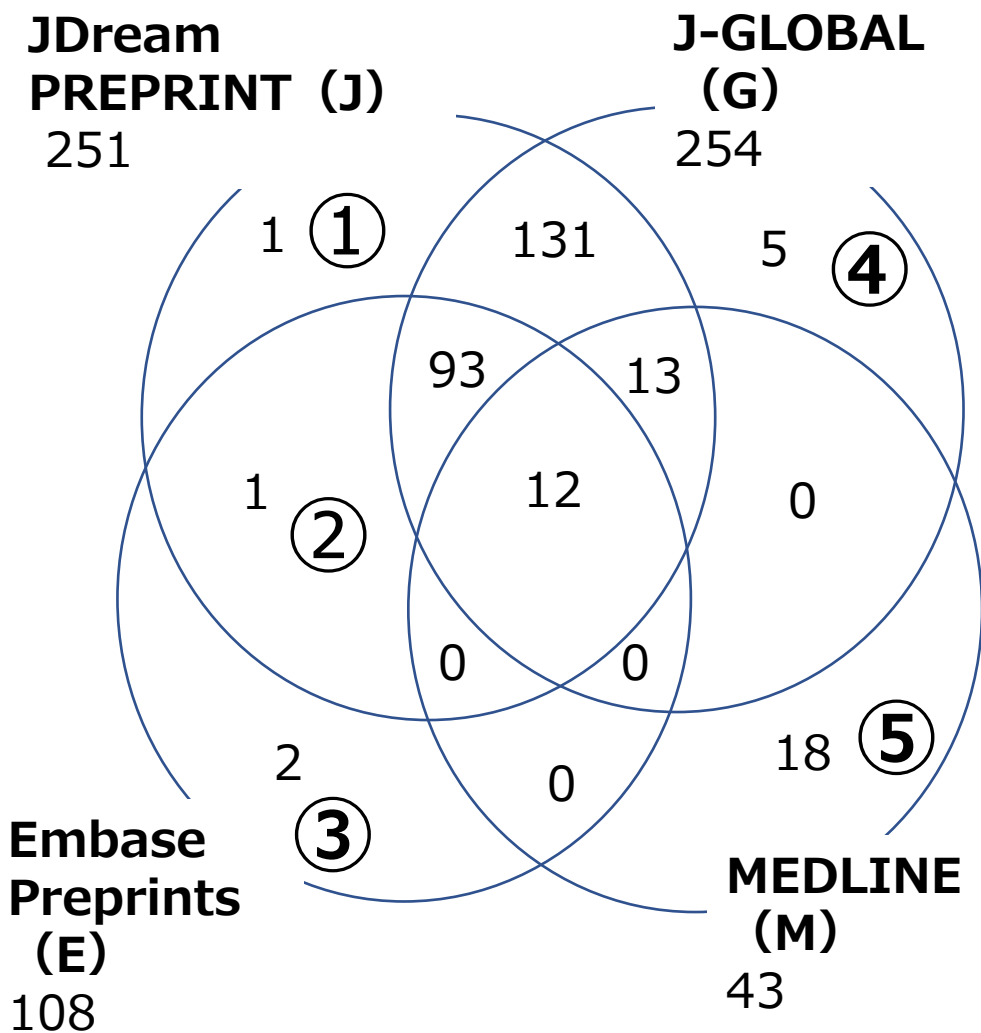
タイトル中「COVID-19」が「-」区切りでなかったためJ-GLOBALにも収録されていたがヒットしなかった。

データベース	英文標題
①JDreamPREPRINTのみの1件 : arXiv	Deep Learning For Classification Of Chest X-Ray Images (Covid 19)
②EmbasePreprintsとJDreamPREPRINTの1件 : medRxiv	Electronic data review, client reminders, and expanded clinic hours for improving cervical cancer screening rates after COVID19 pandemic shutdowns: a multi-component quality improvement program
なお②はEmbasePreprintsのタイトルでは「COVID-19」と記載されていた : 別タイトル	Electronic data review, client reminders, and expanded clinic hours for improving cervical cancer screening rates after COVID-19 pandemic shutdowns: a multi-component quality improvement program

その後の収録 :

- ①②とも、その後もMEDLINEに収録されなかった。
(なお①はarXivなのでEmbasePreprintsの収録対象外である)

COVID-19 ③④⑤その後の確認



③ Embase Preprints (E) のみの2件 : medRxiv
その後、GJで1件ヒットしたが、**もう1件は収録なし**。
Mは2件ともなし。

④ J-GLOBAL (G) のみの5件 : medRxiv
Jには5件とも収録されていたことが判明 (年違いでヒットせず)。
Eにもすべて「年」違いで収録されていることを確認。
Mは5件とも収録なし。

⑤ MEDLINE (M) のみの18件 : medRxiv(6)、
bioRxiv(2)、arXiv(2)、ResearchSquare(8)
その後GJで7件ヒットし5件がmedRxiv でEもヒット、2件はarXiv。
その他、EではmedRxiv1件とbioRxiv1件がヒットしたが、
これら2件はその後もGJで収録なし。
**このほかbioRxivの1件がEに「年」違いで収録されていたが、
同じデータがGJでタイトル違いでヒットしなかったことを確認**。

③ Embase Preprintsのみの2件 : medRxiv

- ✓ J-GLOBALとJDreamPREPRINTで1件は収録を確認したが、1件は収録されなかった。
(下表の2行目)
- ✓ 2件ともMEDLINEにはその後も収録されなかった。

Title	確認結果
A Randomized Trial Comparing Omicron-Containing Boosters with the Original Covid-19 Vaccine mRNA-1273	J-GLOBALとJDreamPREPRINTは収録を確認。 MEDLINEはその後も未収録
Current Status and Emerging Trends of COVID-19-related Studies in Seven 'Tropical Medicine'-entitled Journals	J-GLOBALとJDreamPREPRINT、および、MEDLINEでその後も未収録

④ J-GLOBAL のみの5件 : medRxiv

- ✓ その後JDreamPREPRINTでも全てヒット（両DBに収録済みのデータが更新される場合はJ-GLOBALが先に更新される：先行して出版年が2023年に変更されていた）。
- ✓ 5件ともEmbasePreprintsにも収録されていたが出版年が2023年でなかった。
- ✓ その後もMEDLINEには収録されていなかった。

Notes (英文標題)	確認結果
Evaluation of primary allied healthcare in patients recovering from COVID-19: first results after six months follow-up in a Dutch nationwide prospective cohort study	J-GLOBALとJDreamPREPRINTの違いは 更新タイミングで「年」に更新があったため 。 EmbasePreprintsは「年」違いでヒットしていなかった。
Polygenic Risk Scores for Asthma and Allergic Disease Associate with COVID-19 Severity in 9/11 Responders	
Inflammasome activation and pulmonary viral loads define two distinct clinical outcomes in COVID-19	
Reversal of increases in methadone distribution for opioid use disorder in the US during the COVID-19 pandemic	
Impact of the COVID-19 pandemic on the socioeconomic inequality of health behavior among Japanese adolescents: a two-year-repeated cross-sectional survey	
	MEDLINEはその後も未収録。

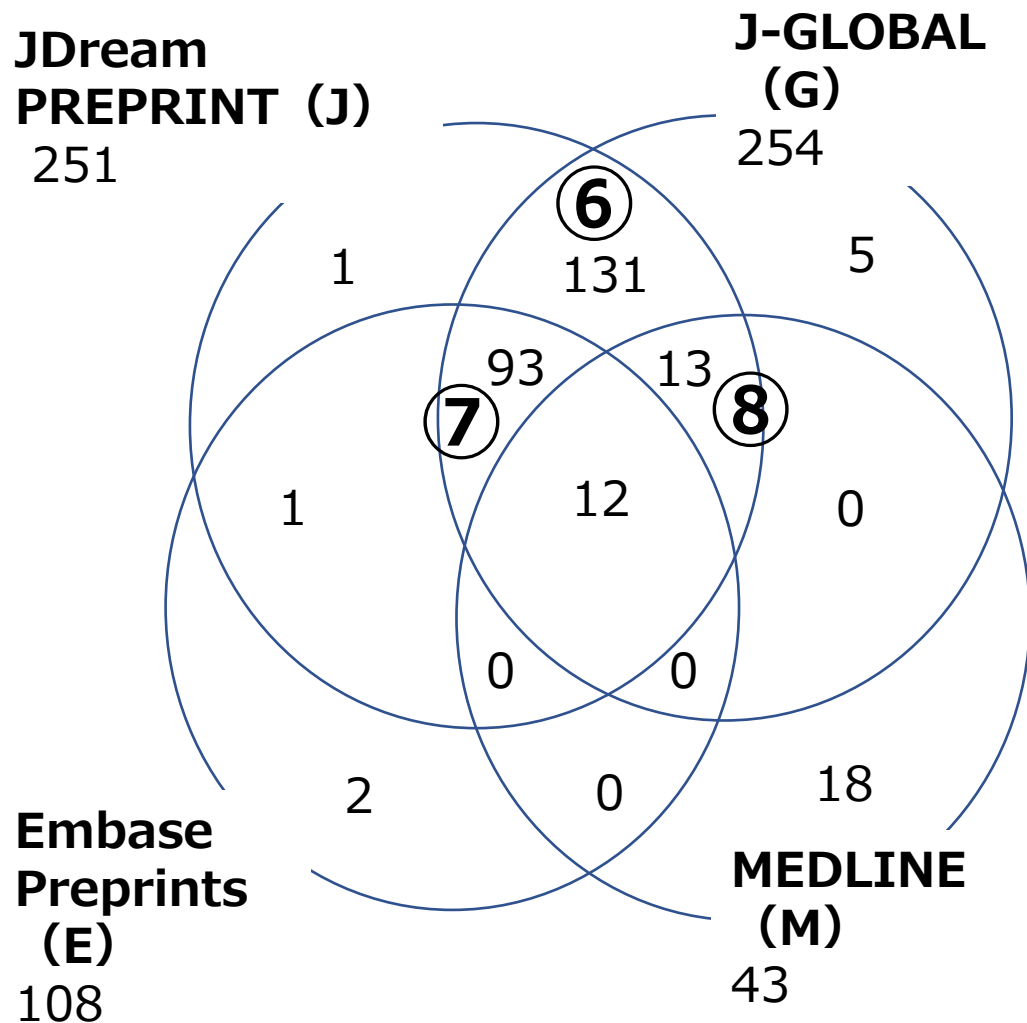
⑤MEDLINEのみの18件：medRxiv(6)、bioRxiv(2)、arXiv(2)、ResearchSquare(8)

- ✓ その後、J-GLOBAL、JDreamPREPRINTで7件ヒット（medRxiv5件、arXiv2件）
EmbasePreprintsでもmedRxivの5件がヒットした。
- ✓ **EmbasePreprintsでヒットしたmedRxivとbioRxivが他2DBではヒットしなかった。**
- ✓ **EmbasePreprintsに「年」違いで収録されていたbioRxivの1件は、他2DBでは別タイトルのためヒットしなかった。**

EmbasePreprintsでヒットし他2DBでヒットしなかった2件、および、年違い・別タイトルの1件

Title	種別	その後の収録
Ischemic stroke after COVID-19 bivalent vaccine administration in patients aged 65 years and older: analysis of nation-wide patient electronic health records in the United States	medRxiv	EmbasePreprints：あり J-GLOBAL・JDreamPREPRINT：なし
Single-cell multi-omic topic embedding reveals cell-type-specific and COVID-19 severity-related immune signatures	bioRxiv	EmbasePreprints：あり J-GLOBAL・JDreamPREPRINT：なし
COVID-19 -associated pulmonary aspergillosis in immunocompetent patients: A virtual patient cohort study 別タイトル ： COVID -associated pulmonary aspergillosis in immunocompetent patients	bioRxiv	EmbasePreprints：「年」違い J-GLOBAL・JDreamPREPRINT：別タイトル

COVID-19 検索結果件数



⑥ **131** : medRxiv(61)、bioRxiv(18)、arXiv(52)

その後、Eで48件ヒットmedRxiv(35)、bioRxiv(13)。

arXivを除いた残り**31件が「年」違いで収録あり。**

(ここでのE収録はmedRxivとbioRxivはGJと同じ)

Mで追加の収録なし。

⑦ **93** : medRxiv(93)

Mで追加の収録なし

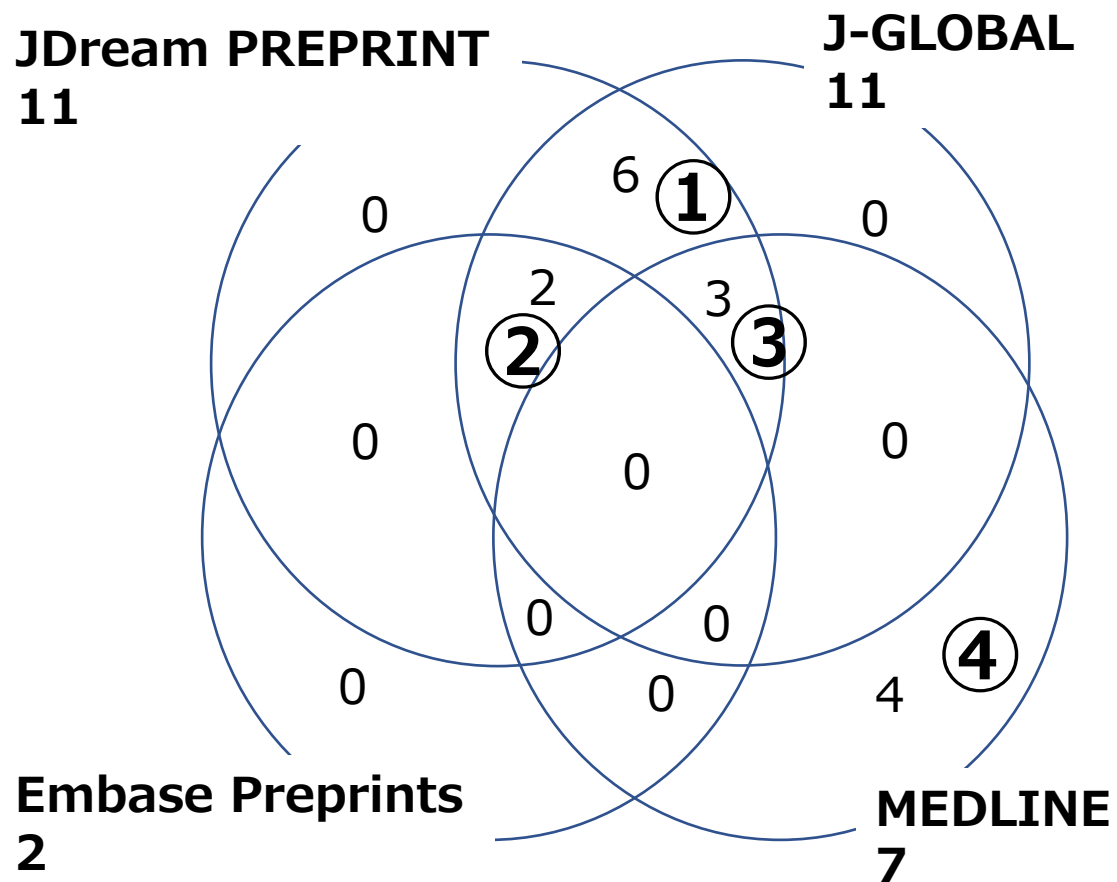
⑧ **13** : medRxiv(11)、bioRxiv(2)

その後、Eで8件ヒット、medRxiv(6)、bioRxiv(2)。

残り5件は「年」違いで収録あり。

(ここでのE収録もmedRxivとbioRxivはGJと同じ)

Hypertension



データベース	件数
JDream PREPRINT	11
J-GLOBAL	11
Embase Preprints	2
MEDLINE	7

①6 :
Eで4件ヒット。**残り2件は「年」違いで収録されていることを確認。**
その後もMでヒットしない。

②2 :
その後もMでヒットしない。

③3 :
その後、Eで3件ともヒット。

④4 :
その後、GJEともに共通で2件ヒット : medRxivとbioRxiv。
残り2件はResearchSquareなので収録対象外。

結果

- JDreamPREPRINTとJ-GLOBALでは、ハイフン「-」の検索仕様の違いによる検索結果の違いを認めた。
- JDreamPREPRINTとJ-GLOBALは同じデータと思われたが、両データベースに更新タイミングの差があることを確認した。
- データベースへの収録にタイムラグのあることを確認した。
- データベースによって「出版年」や「タイトル」が異なるものがあった。
- JDreamPREPRINTとJ-GLOBALで収録対象だが時間が経過しても収録されないデータが存在した（2023年6月時点）。
- MEDLINEに収録されないプレプリントは、収録要件であるNIHの研究助成がなされていないものと思われた。

考察

- ハイフン有無などの検索仕様に対する注意が必要なことを確認できた。
- データベースごとに収録タイミングの違いのあることを認識する必要がある。
- 「年」の違いはプレプリントの性質上やむを得ない部分もあるが、プレプリントのバージョン情報、また、掲載や更新時期を認識できる仕組み作りなど利用面での工夫がデータベース側にあるとよいと考えた。
関連して、プレプリントのバージョンなどの情報がデータベースと連動・協調することが望ましいと考えた。
- データベースごとに「年」や「タイトル」が違う要因として、採用したバージョンやデータ領域の違いがあるものと思われたため、データベース構築におけるデータの内容の提示（透明性の向上）が望ましいと考えた。
- JDreamPREPRINTとJ-GLOBALで収録対象だが時間が経過しても収録されないデータが存在した点については、タイムラグの短縮化など今後のデータベース構築の改善につながることを希望する。

最後に

今回の検証にあたり、データベース作成・提供機関のJST様、ジー・サーチ様など関係の皆様のご協力をいただきましたことに感謝申し上げます。

ご清聴をありがとうございました。