Brief Report: Research Publications Related to the Corona Virus from PUBMED search July 15, 2020

We have been downloading abstract record sets from the PubMed database relating to corona virus every 2 weeks, starting March 25. See: http://www.techminingforglobalgood.org/open-covid-19-research-for-analysis/.

For the first several datasets, we generated basic reports, summarizing rudimentary analyses to introduce the data for potentially deeper probes. Those reports remain available; they variously address:

- The data & background on the searches
- Looking at the data in *VantagePoint* desktop text analysis software [www.theVantagePoint.com]
- Research Trend
- Notes on players participating strongly in that research (countries, organizations, researchers)
- Changing topical emphases
- Exploratory analysis (trying one novel empirical effort each time) – e.g., for May 21, we compared topic emphases through May 6, 2020 to emphases from then forward.

Here, we offer a more concise overview with a new exploration, but first address issues in searching and downloading the records. Starting with May 20, we encountered limitations in PubMed downloading.

Search Essentials

We apply the search query, but apply it to the legacy version of the PubMed database [by shifting from "PMC' via pull-down to "PubMed"][1]

- "COVID-19" OR Coronavirus OR "Corona virus" OR "2019-nCoV" OR "SARS-CoV" OR "MERS-CoV" OR "Severe Acute Respiratory Syndrome" OR "Middle East Respiratory Syndrome"[2]

The interface now limits downloads to the first 10,000 records, so we ran a combination of partial searches to recombine in *VantagePoint* on desktop. We use the core query: (search query as just above), plus additional constraints or terms to divide into datasets <=10000.

- See prior reports for more information on our searching. This time we just download updated 2020 publications in legacy PubMed, to be combined with prior year results on our desktop.
- Beware that the interface resets to all years upon changing the query
- Further restricted to 2020-21, as it had 33569 records [2 of those dated 2021]
    - (search query as above) with added terms to restrict further:
        - AND coronavirus = 20973
            - AND viral = 13609
                - AND epidemiology = 5963 saved
                - NOT epidemiology = 7646 saved
            - NOT viral = 7364 saved
        - NOT coronavirus = 12596
            - NOT (study OR research) = 8392 saved
            - AND (study OR research) = 4204 saved

[1] We use the legacy, peer-reviewed journal papers for consistency. PMC includes preprints, many lacking metadata indexing and some full text that does not combine well with analyses of abstract records.

[2] In MEDLINE detail: "COVID-19"[All Fields] OR ("coronavirus"[MeSH Terms] OR "coronavirus"[All Fields]) OR "Corona virus"[All Fields] OR "2019-nCoV"[All Fields] OR "SARS-CoV"[All Fields] OR "MERS-CoV"[All Fields] OR "Severe Acute Respiratory Syndrome"[All Fields] OR "Middle East Respiratory Syndrome"[All Fields].

- Downloaded by choosing: Save; Selection: all results; Format: PubMed; Create File [downloads to desktop computer]
- Imported using the PubMed configuration file; removed duplicates; and thereby consolidated these in one *VantagePoint* file = 33569 records.
- Merged files with July 1 dataset that included prior year results; removed duplicates using PubMed ID to get 52032 [slightly differing from the current PubMed count of 52020].

**Prior year dataset counts indicate high stability over these repeat searches (i.e., only 5 new records on July 1 dated pre-2020). So, in future searches, one could just add for 2020 on and consolidate with prior year results (removing duplicates).

Results

Series of Bi-weekly searches:

- March 25, 2020: Initial dataset - 19538 records.
- April 8, 2020: 21314
- April 22, 2020: 24479
- May 6, 2020: 28264
- May 21, 2020: 34181
- June 3, 2020: 38183
- June 17, 2020: 43254
- July 1, 2020: 47607
- July 15, 2020: 52032

Explosive growth in COVID-19 research publication this year continues, but the number of new publications each two weeks is steadying.

As mentioned, we elect not to redo the rudimentary research profiling presented in prior biweekly reports. Those remain available to peruse research participation, topical emphases, and research trend. For those of you wanting to answer "**who, where, when, what?"** questions, we make the data available in *VantagePoint* format, along with free trial use of the software to explore your interests.

We have spotlighted different special analyses in the biweekly reports, to stimulate consideration of other analytical inquiries that might generate research intelligence of value in countering COVID-19. Let us know if you have a particular interest. Prior analyses include:

- March 25 and several following biweekly reports gave screenshots of the dataset in *VantagePoint* and showed the explosive growth trend for this literature. They also tallied statistics on research participation, publications, and topical emphases.
- March 25 additionally mapped **collaboration among countries.**
- April 8 broke out **top MeSH (Medical Subject Headings) Descriptors by recent month**
- April 22 **clustered topics**
- May 6 compared MeSH Qualifiers pre-2020 vs. 2020 to date.
- May 21 spotlighted **topics (title phrases) growing most strongly in the prior 2 weeks**.
- June 3 identified leading author organizations publishing in this literature in 2020 and MeSH Qualifier Topics escalating in coverage from Jan.-Apr, 2020 to May, 2020 forward.
- June 17 and July 1 explore **prevalence of terms associated with different approaches** to addressing the pandemic.

This report experiments with "what's really new?" analyses. We make a separate dataset of 4425 abstract records new in PubMed since July 1. Were we to separate by publication dates, the dataset would change

somewhat. The 4425 include some 228 dated "2020" or such (e.g., 2020 Jan-Dec) and 1188 dated pre-July. Here, we analyze these records added to PubMed since July 1.

We briefly address basic "**who, where, when, what?"** questions for these recent couple weeks.

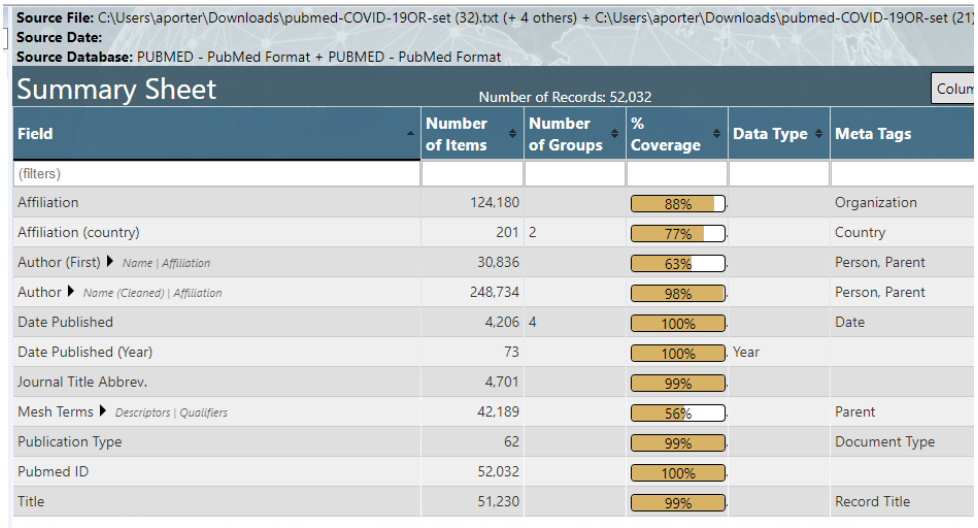
- Who?
  - Top 8 organizations (not fully cleaned): Huazhong Univ of Sci & Tech; Univ of Toronto; Univ of California; Oxford Univ; Mass. General Hosp; Tehran Univ of Med Sci; Univ of Milan; Mayo Clinic
- Where?
  - Iran and Brazil nudge into the Top 12 for this period
- When? Current 2 weeks – we would be interested in feedback on whether "what's new" is of interest. And if so, what period is best (e.g., perhaps 1 month)?
- What?
  - MeSH coverage is too low here – 7%
  - So, we revisit the "Emergent Topics" derived in the full July 1 data analyses. See Appendix for "how to," as well as a screenshot of the 4425 dataset in *VantagePoint.*
  - Table 1 shows the relative 'acceleration' of research on these hot topics (within a blazing domain). The ratio shows the number of records including one of the high-loading terms of the Emergent Topics in the new 4425 (~July 2-15) dataset vs. the number in May-June, 2020. The overall ratio of records for the two periods is 4425/21405 = 20.7%; the hi-lited ones grew faster than the dataset growth for these recent periods. The Topics are sorted in order of relative gain. [That most of the "emergent" topics show growth rate above the overall COVID-19 dataset rate supports our intent to distinguish particularly hot topics, even in such a booming research domain.

	4425	20.7%
Combo terms NLP (1) NLP Refined : COVID-7-1 ETs (2)	**# Records**	**Ratio of July 2-15 to May-June**
Antibodies (IgM, IgG)	58	56.3%
Surgery	151	53.4%
Surveys and Questionnaires (health knowledge, attitudes, practice)	103	43.6%
Anticoagulants (thrombosis)	110	39.6%
Intensive Care Unit (ICU)	331	38.6%
Mental health	136	37.6%
T cells	23	37.1%
Cancer	147	35.3%
Computerized Tomography (CT)	204	35.1%
Molecular docking	34	34.7%
Arrhythmias	58	31.4%
Study protocol (e.g., randomization)	87	31.2%
C-reactive protein (e.g., D-dimer)	170	31.1%

Angiotensin-converting enzyme inhibitors (ACE2)	139	30.5%
Embase	131	30.0%
Diabetes	167	28.8%
Education, Medical	40	27.6%
Symptoms (e.g., fever)	169	27.1%
Olfactory disorders (Anosmia)	46	26.4%
Neurological manifestations	36	24.0%
Personal protective equipment	177	23.0%
Ophthalmology	20	22.0%
Healthcare disparities (inequities)	79	21.5%
Hydroxycholoroquine (HCQ)	114	21.3%
Ritonavir/Lopiinavir	38	18.5%
Infectious Disease Transmission, Vertical (pregnaancy)	26	15.8%
Orthopedics	16	12.5%
Head and Neck Neoplasms	15	10.4%
Patient Care Teaming	58	9.2%

APPENDIX

The July 15 "minus" July 1 Dataset = 4425 Records



Source File: C:\Users\aporter\Downloads\pubmed-COVID-19OR-set (32).txt (+ 4 others) + C:\Users\aporter\Downloads\pubmed-COVID-19OR-set (21)
Source Date:
Source Database: PUBMED - PubMed Format + PUBMED - PubMed Format

Summary Sheet — Number of Records: 52,032

Field	Number of Items	Number of Groups	% Coverage	Data Type	Meta Tags
(filters)					
Affiliation	124,180		88%		Organization
Affiliation (country)	201	2	77%		Country
Author (First) ▸ *Name | Affiliation*	30,836		63%		Person, Parent
Author ▸ *Name (Cleaned) | Affiliation*	248,734		98%		Person, Parent
Date Published	4,206	4	100%		Date
Date Published (Year)	73		100%	Year	
Journal Title Abbrev.	4,701		99%		
Mesh Terms ▸ *Descriptors | Qualifiers*	42,189		56%		Parent
Publication Type	62		99%		Document Type
Pubmed ID	52,032		100%		
Title	51,230		99%		Record Title

"How-to" for Calculating the Prevalence of Emerging Topics in this recent 2 weeks:

- Imported Abstract (Abs) NLP (Natural Language Processing) noun phrases; applied Refine NLP script to remove stopwords and consolidate term variants
- Imported Title (Ti) NLP phrases and applied Refine NLP
- Combined Abs + Ti NLP phrases (refined) + MeSH Descriptors (despite low coverage),

- Applied COVID Abs terms ListCleanup-8-3 thesaurus to Combo terms NLP
- Applied Refine NLP to Combo Terms.
- Took the 1347 Emergent Terms from the July 1 analysis as a "My Keywords" list (COVID-7-1 ETs) and extracted those from the 4425 record set's 60,334 Combo Terms NLP (1) NLP Refined. Found 1299 in the 4425 set.
- Applied a thesaurus -- PubMed E Topics-7-1-2020 -- (each being a group of Emergent Terms based on PCA – Principal Components Analysis – in *VantagePoint* run on the 7-1 dataset).