Research Publications Related to the Corona Virus from PUBMED

The Data

The U.S. Government is helping to provide access to research related to Covid-19. A number of resources are being made available, including:

- https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/
- Follow the link to here: https://pages.semanticscholar.org/coronavirus-research
- Follow the link to "PubMed's PMC open access corpus using this query (COVID-19 and coronavirus research)"

These resources provide various agglomerations of research drawn from multiple resources. Here, we take a search query [available at "query" link above][1] and applied it to the legacy version of the PubMed database:

- "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"

We prefer using the legacy interface because it provides MEDLINE indexed abstract records of published research appearing in scholarly journals. This offers a good degree of comparability, whereas mining other of the compilations demands considerable attention to what is being mixed together. The PubMed PMC open access interface includes preprints, some records lack metadata, and some records are full text – so it is potentially hard to gauge how complete or representative such a dataset is.

Our initial dataset of 19,538 records is from a search run on March 25, 2020. We intend to update this periodically as feasible. Here is the April 8, 2020, update now yielding **21,314** records. We will keep the simple research profiles from prior searches available, should you want to revisit. The one based on the March 25 search contains similar content as this one, less ??

We make this latest dataset available as a "vpt" format file for further text analyses using *VantagePoint* software [www.theVantagePoint.com]. We also make that software available free on a 60-day trial basis for researchers and analysts who want to investigate particular treatments or biomarkers, explore associations, etc.

Perspective on the Covid-19-related Data

Figure 1 offers a screenshot of the file summary sheet in *VantagePoint.* A quick tour of the dataset provides some perspective on the biomedical research activity.

- Breadth of **Research Participation** is extensive:
  - Authors or co-authors of the papers reside in 143 countries
  - 26 countries host researchers publishing 100 or more of these papers
  - Of these 19,538 publications, the only 2 countries authoring more than 1000 of them are the US (4228) and China (4115)

[1] 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].

- Some surprises to an outsider – e.g., Saudi Arabia shows as #13 with 419 papers.

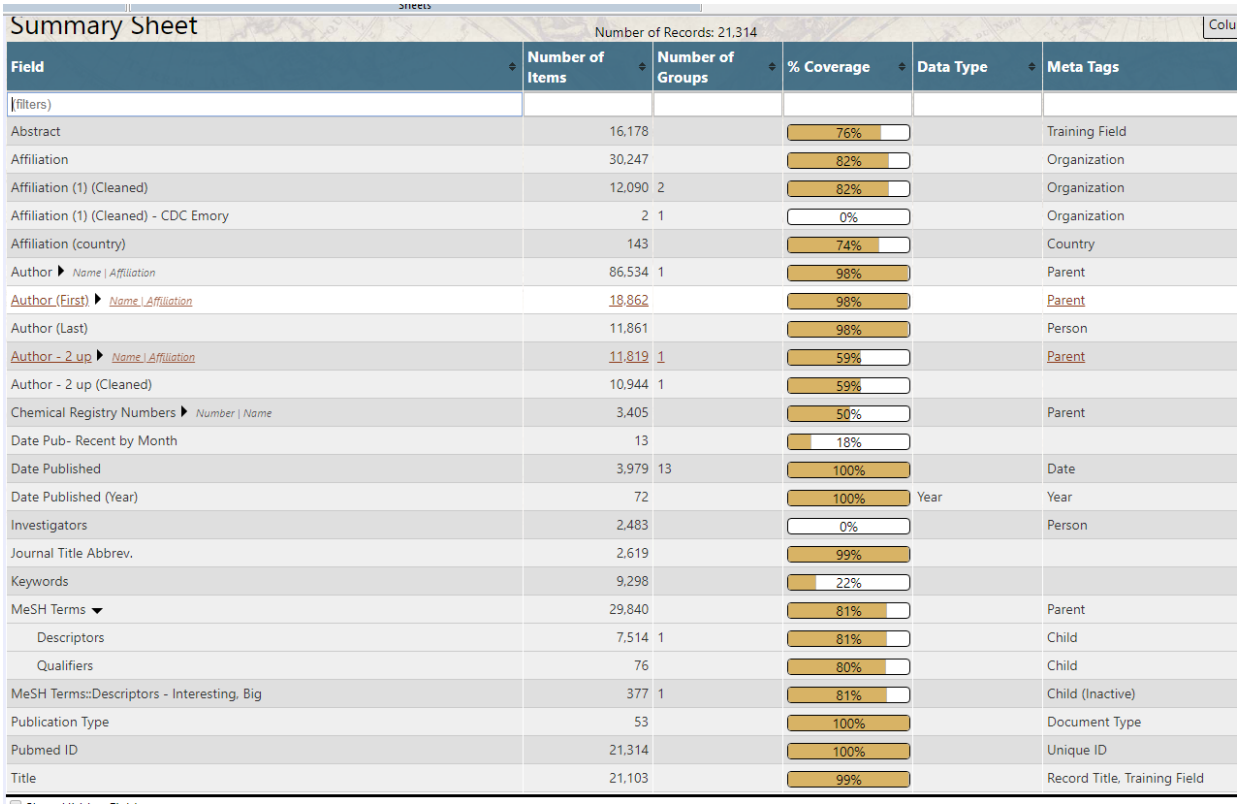
- **Publication Types**: 18846 of these 21314 publications are journal articles (including reviews). For some analyses, one might want to focus on those.
  - Only 55 are identified as clinical trials.

- **Topical Content**:
  - Medical Subject Heading Terms (MeSH) appear for some 81% of the records – a few tallies of MeSH descriptors:
    - About half of the papers address human subjects (9238) and half, animals (9718)
    - Animals studied are most prominently: mice (2793 of the papers), swine (1502), chickens (1080), and cattle (676)
  - Analyses of intersections represented can help focus – e.g., papers addressing "coronavirus infections" (4765) appear almost equally in records with each of the three leading animal subjects – mice (683), swine (659), and chickens (624).
  - Analyses of MeSH qualifiers could also serve to discern pockets of research of special interest to a researcher – e.g., many studies address genetics (6082) and immunology (4536), prevention & control (2884), transmission (1705), drug effects (1145), and so forth.
  - Further analyses of the topical content (e.g., from titles and abstract noun phrases) holds promise to enable focus on particular attributes and their intersections in subsets of studies. *VantagePoint* facilitates such analyses.

- Quick scans of the **top researchers**[2] show
  - 447 organizations affiliated with 10 or more of the 19538 papers
  - 16 organizations affiliated with 100 or more papers; of which the top two are University of Hong Kong and Chinese Academy of Sciences.
  - Top authors (110 or more of these papers) are Stephen Stohlman, Kwok Yung Yuen, Steven Baric, Susan Weiss, Ralph Luis Enjuanes, Stanley Perlman, Michael Lai, and Kathryn Holmes [See Ftn. 2]

[2] PubMed data on Authors and Author Affiliations tallied here are partly cleaned in *VantagePoint* using tailored fuzzy match, list cleanup routines tuned to authors and their affiliations. Full, detailed tuning by humans was not done.

Figure 1. *VantagePoint* Summary View of the Dataset



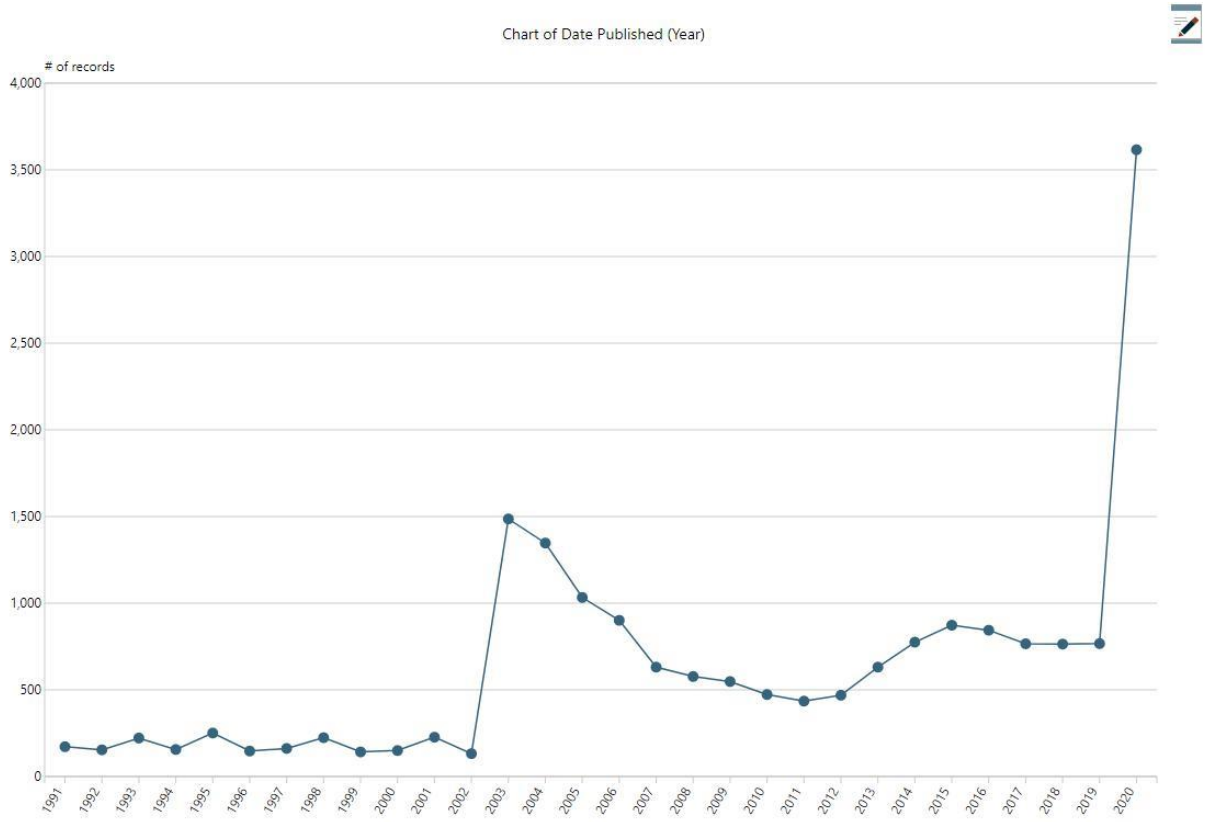
Summary Sheet

Number of Records: 21,314

Field	Number of Items	Number of Groups	% Coverage	Data Type	Meta Tags
(filters)					
Abstract	16,178		76%		Training Field
Affiliation	30,247		82%		Organization
Affiliation (1) (Cleaned)	12,090	2	82%		Organization
Affiliation (1) (Cleaned) - CDC Emory	2	1	0%		Organization
Affiliation (country)	143		74%		Country
Author ▸ *Name | Affiliation*	86,534	1	98%		Parent
Author (First) ▸ *Name | Affiliation*	18,862		98%		Parent
Author (Last)	11,861		98%		Person
Author - 2 up ▸ *Name | Affiliation*	11,819	1	59%		Parent
Author - 2 up (Cleaned)	10,944	1	59%		
Chemical Registry Numbers ▸ *Number | Name*	3,405		50%		Parent
Date Pub- Recent by Month	13		18%		
Date Published	3,979	13	100%		Date
Date Published (Year)	72		100%	Year	Year
Investigators	2,483		0%		Person
Journal Title Abbrev.	2,619		99%		
Keywords	9,298		22%		
MeSH Terms ▾	29,840		81%		Parent
Descriptors	7,514	1	81%		Child
Qualifiers	76		80%		Child
MeSH Terms::Descriptors - Interesting, Big	377	1	81%		Child (Inactive)
Publication Type	53		100%		Document Type
Pubmed ID	21,314		100%		Unique ID
Title	21,103		99%		Record Title, Training Field

- The **Research Trend** is informative:
  - From the first paper in 1949, we see fewer than 10 papers/year until 1962.
  - We first see over 100 papers in a year in 1981 (40 years ago!).
  - Then, over 1000 papers per year first in 2003.
  - Figure 2 shows the publication trend over the last 20 years – note the explosion of attention in 2003 and, again, now (more so) in 2020 (and that's just in the beginning of this year).
  - Messages herein include that research attention is "exploding" nowadays, but corona-related viruses have been studied for decades.

Figure 2. Trend in Corona Virus Related Publications from PUBMED



- Exploring research interests; this dataset can help answer basic "**who, what, where, when?"** questions. Juxtaposing the available fields of data can point to research activity of particular concern:
    - For instance, a matrix of **Authors (or Organizations) X Topics** could point to hotbeds of research on particular corona virus subjects.
    - A matrix of **Topics X Years** can help gauge the emergence of various topics. Figure 3 shows # of PubMed-indexed papers over time. Col. B gives the total for all years (e.g., 2924 for SARS Virus); Columns D-P show the counts for recent months.[3] These are the MeSH Descriptors with 900 or more publications associated.
        - The recent explosion of research is notable – The second row from the top gives monthly totals. E.g., March, 2020, swamps any other month with 1935 publications, up from 501 in February, and 98 in January.
        - Note also that for most of these predominant topics, most of the publications precede the dates shown – e.g., for SARS virus, all but 60 of 2924 papers addressing it are earlier.
        - Of possible interest, some topics show minimal recent interest – e.g., "murine hepatitis virus" – might some of these warrant revisiting?
        - Other topics show sharp recent publication bursts – e.g., "China."
    - Or, one might inspect a matrix of **Topics X Topics** to pursue intersecting factors (e.g., viral antibodies together with assays; viral RNA with cell lines).
    - Another matrix could counterpose active researchers in particular countries who have recently published on a topic of special interest (e.g., American researchers engaging viral RNA re: coronaviruses).

- *VantagePoint's* import filter recognizes some 50 additional fields in these MEDLINE records, enabling many specialized explorations.

Figure 3. Top MeSH Descriptors by Frequency in Recent Months

		MeSH Terms::Descriptors - Interesting	1	2	3	4	5	6	7	8	9	10	11	12	13
		# Recor	60	53	52	51	44	52	65	57	69	98	501	1935	948
	# Records	Date p	2019(Apr)	2019(May)	2019(June	2019(July)	2019(Auf)	2019(Sep)	2019(Oct)	2019(Nov)	2019(Dec)	2020(Jan)	2020(Feb)	2020(Mar)	2020-(Apr+)
1	4493	Severe Acute Respiratory Syndrome	4	3	2	4			4	1	1	5	9	26	7
2	2924	SARS Virus	3	3		1	1		1		1	5	7	31	7
3	2793	Mice	2	5	3	3			3	1	1	3	1	3	1
4	2217	Disease Outbreaks	2	5	4	2	2	2	5	2	1	13	43	72	27
5	2101	Murine hepatitis virus	1	1										1	
6	2001	Antibodies, Viral	8	4	2	2	4	2	3	3	2	2	1	9	2
7	1905	Coronavirus	5	5	7	1	1	2	3	1	5	12	24	37	18
8	1719	RNA, Viral	5	1	1	1	5	1	2	1	3	2	2	2	4
9	1655	Cell Line	3	4	3		1	1	2	3	2		2	3	1
10	1502	Swine	17	6	8	6	6	7	9	7	3	2	2	1	
11	1501	Infectious bronchitis virus	8	5	1	7	2	3	6	5	1			1	
12	1487	Middle Aged	3	5	6	3	2	4	1	2		2	15	36	11
13	1328	Coronaviridae								1					
14	1322	Virus Replication	7	2	3		3	1	3	1	3	3	3	4	1
15	1310	Spike Glycoprotein, Coronavirus	2	5	6	4	1	1	3	3	2	2	2	12	6
16	1290	Viral Proteins	5	1		2	2	3	1		3	2	3	2	2
17	1216	Amino Acid Sequence	2		1	1		2	2		1			3	3
18	1170	Viral Envelope Proteins		2	1							1		5	3
19	1130	China	4	2	1	2	2	2	1	3	2	12	52	117	32
20	1121	Base Sequence	1						1	1					1
21	1097	Phylogeny	10	4	4	3	3	3	8	1	2	1	8	11	10
22	1080	Chickens	8	3	2	4	2	2	6	5	2			1	
23	1066	Viral Vaccines	5	3		5	1	3	4	3		2	4	12	2
24	988	Aged	3	3	5	3		4	1	2		1	8	30	6
25	986	Middle East Resp Syndr Coronavirus	6	8	7	6	5	5	4	3	4	4	3	16	4
26	921	Membrane Glycoproteins													1
27	904	Poultry Diseases	8	2	2	5	2	2	5	5	2				
28	900	Respiratory Tract Infections	4	7	5	2	3	3		1	2	1	4	1	1

[3] Date accuracy at the monthly level is imperfect.