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. We make this 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 140 countries
  - 25 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 (3915) and China (3672)
  - Some surprises to an outsider – e.g., Saudi Arabia shows as #12 with 407 papers.

- **Publication Types**: 17601 of these 19,538 publications are journal articles (including reviews). For some analyses, one might want to focus on those.

[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].

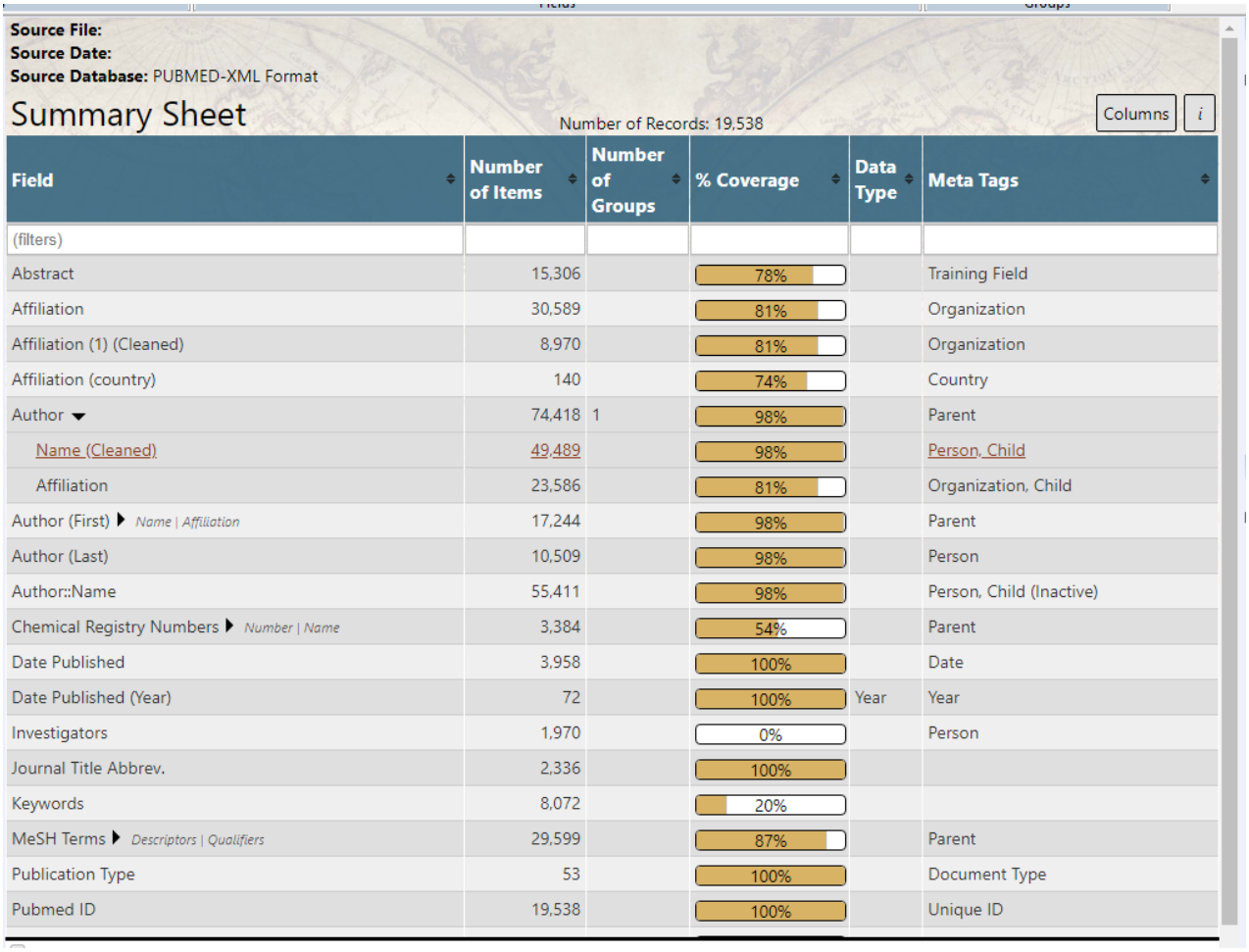
- Only 55 are identified as clinical trials

- **Topical Content**:
  - Medical Subject Heading Terms (MeSH) appear for some 87% of the records – a few tallies of MeSH descriptors:
    - About half of the papers address human subjects (9058) and half, animals (9693)
    - Animals studied are most prominently: mice (2790 of the papers), swine (1498), chickens (1080), and cattle (676)
  - Analyses of intersections represented can help focus – e.g., paper addressing "coronavirus infections" (4581) appear almost equally in records with each of the three leading animal subjects – mice (681), swine (655), 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 (6060) and immunology (4517), prevention & control (2821), transmission (1656), drug effects (1136), 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
  - 408 organizations affiliated with 10 or more of the 19538 papers
  - 15 organizations affiliated with 100 or more papers; of which the top two are University of Hong Kong and Chinese University of Hong Kong. Of the top 15, 5 are in the US; 4 in China; 3 in Hong Kong; and 3 elsewhere (Saudi Arabia, The Netherlands; Singapore).
  - Top authors (150 or more of these papers) are Yung Yuen Kwok, Stanley Perlman, Ralph Steven Baric, Luis Enjuanes, Stephen Stohlman, and Susan Weiss. Interestingly, of the top 16 (over 100 papers here), 13 have been publishing corona virus related research since 2003; 11 already have papers out in 2020.

[2] PubMed data on Authors and Author Affiliations (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



Source File:
Source Date:
Source Database: PUBMED-XML Format

Summary Sheet

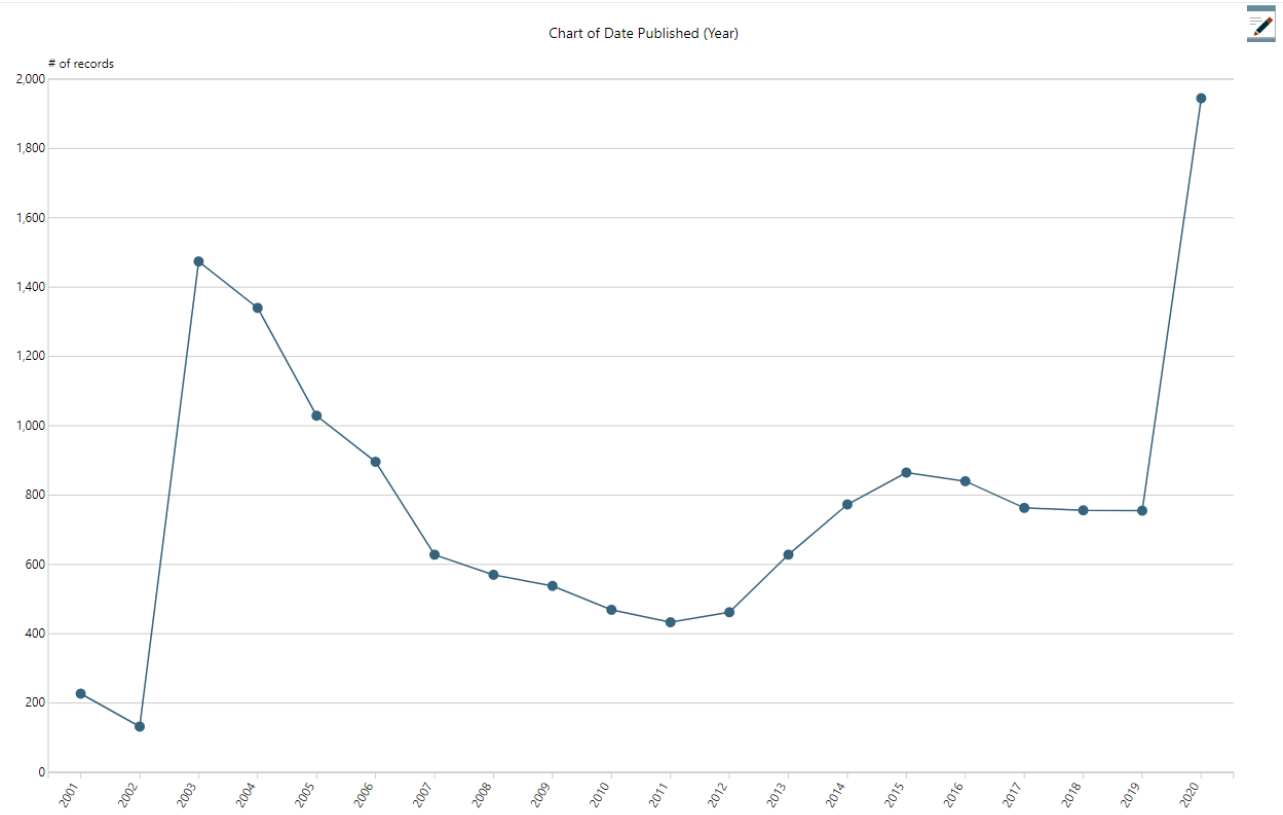
Number of Records: 19,538

Columns | i

Field	Number of Items	Number of Groups	% Coverage	Data Type	Meta Tags
(filters)					
Abstract	15,306		78%		Training Field
Affiliation	30,589		81%		Organization
Affiliation (1) (Cleaned)	8,970		81%		Organization
Affiliation (country)	140		74%		Country
Author	74,418	1	98%		Parent
Name (Cleaned)	49,489		98%		Person, Child
Affiliation	23,586		81%		Organization, Child
Author (First) *Name | Affiliation*	17,244		98%		Parent
Author (Last)	10,509		98%		Person
Author::Name	55,411		98%		Person, Child (Inactive)
Chemical Registry Numbers *Number | Name*	3,384		54%		Parent
Date Published	3,958		100%		Date
Date Published (Year)	72		100%	Year	Year
Investigators	1,970		0%		Person
Journal Title Abbrev.	2,336		100%		
Keywords	8,072		20%		
MeSH Terms *Descriptors | Qualifiers*	29,599		87%		Parent
Publication Type	53		100%		Document Type
Pubmed ID	19,538		100%		Unique ID

- 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** could help gauge the emergence of various topics
    - 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).

- **Collaboration** patterns lend insight to the social networking and the nature of a corona virus research community. Figure 3 shows the extent of co-authoring among countries – i.e., how frequently a paper shows author affiliations from two countries together. Node size reflects rough number of corona virus related papers. The threshold for links to show in this map is fairly high.
    - For instance, the map shows the number of co-authored papers of the two leading countries with their top partners. So, China—US have 267 such papers, but that's only ~7% of each's publications here.

- We see one network of European countries plus Egypt and Saudi Arabia, extending to the US.

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

Figure 3. Collaboration among Countries