Research Publications Related to the Corona Virus from PUBMED Update April 22, 2020

The Data

Here, we take a search query [available at "query" link below][1] and applied it to the legacy[2] 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"

At the PubMed search interface, entering the above search query, we

- "Send to" – pull down menu and select "File."
- Pull down menu for "Format" – we choose XML
- For "Sort by" we leave as "Most Recent"
- Click "Create File"
- Download the 324 MB file (4/22/2020 version) taking ~7 minutes

March 25, 2020: Initial dataset - 19,538 records.
April 8, 2020: Update yielded 21,314 records.
April 22, 2020: Update yielded **24,479** records.

We will keep the simple research profiles from prior searches available, should you want to revisit. The ones based on the prior searches contain similar content as this one, with some variations.

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 152 countries
  - 28 countries host researchers publishing 100 or more of these papers

[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].

[2] 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.

- Of these 24,479 publications, the only 2 countries authoring more than 1300 of them are the US (4857) and China (4752)
- Surprising to an outsider – e.g., Saudi Arabia shows as #13 with 443 papers.

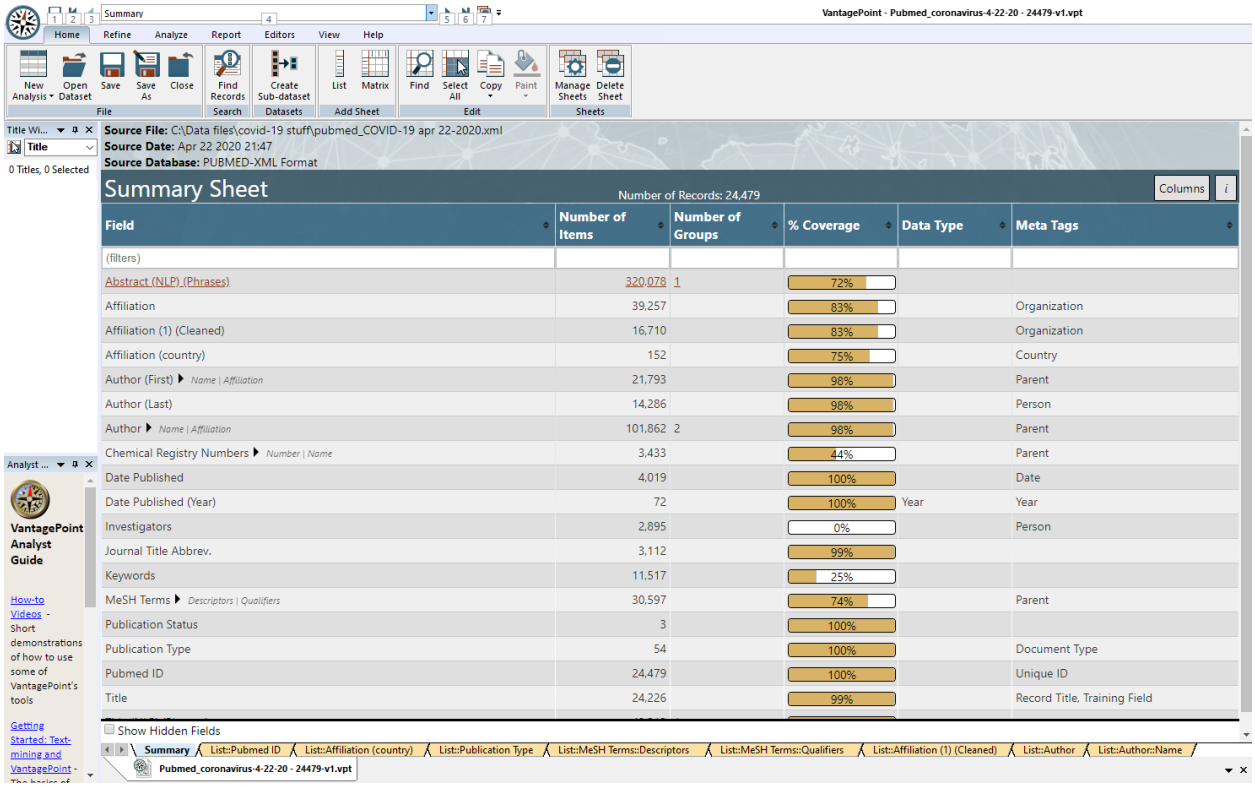
- **Publication Types**: 20967 of these 24479 publications are journal articles (including reviews). For some analyses, one might want to focus on those.
  - Only 55 are identified as clinical trials.

- **Topical Content**:
  - Medical Subject Heading Terms (MeSH) appear for some 74% of the records – a few tallies of MeSH descriptors:
    - About half of the papers address human subjects (9964) and half, animals (9803)
    - Animals studied are most prominently: mice (2801 of the papers), swine (1513), chickens (1085), and cattle (677)
  - Analyses of intersections represented can help focus – e.g., papers addressing "coronavirus infections" (5552) appear almost equally in records with each of the three leading animal subjects – mice (691), swine (669), and chickens (629).
  - Analyses of MeSH qualifiers could also serve to discern pockets of research of special interest to a researcher – e.g., many studies address genetics (6168) and immunology (4596), prevention & control (3110), transmission (1859), drug effects (1180), 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**[3] show??
  - 519 organizations affiliated with 10 or more of the 24479 papers
  - 17 organizations affiliated with 100 or more papers; of which the top two are University of Hong Kong and Chinese Academy of Sciences.

[3] PubMed data on 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.

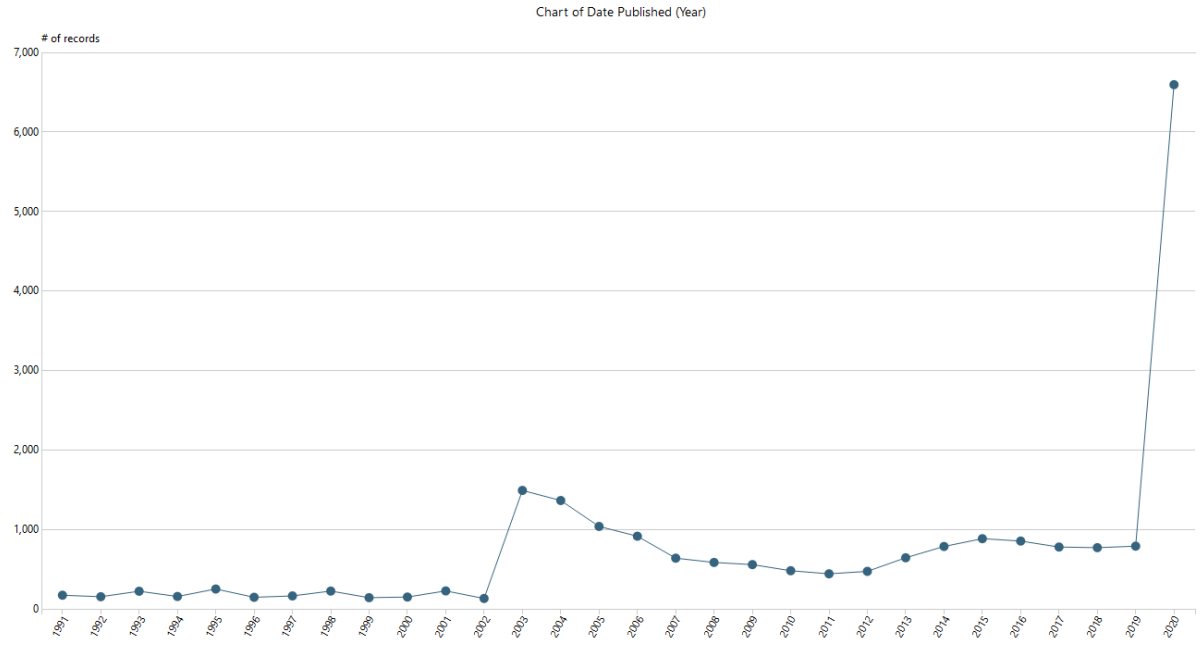
Figure 1. *VantagePoint* Summary View of the Dataset



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 past 30 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



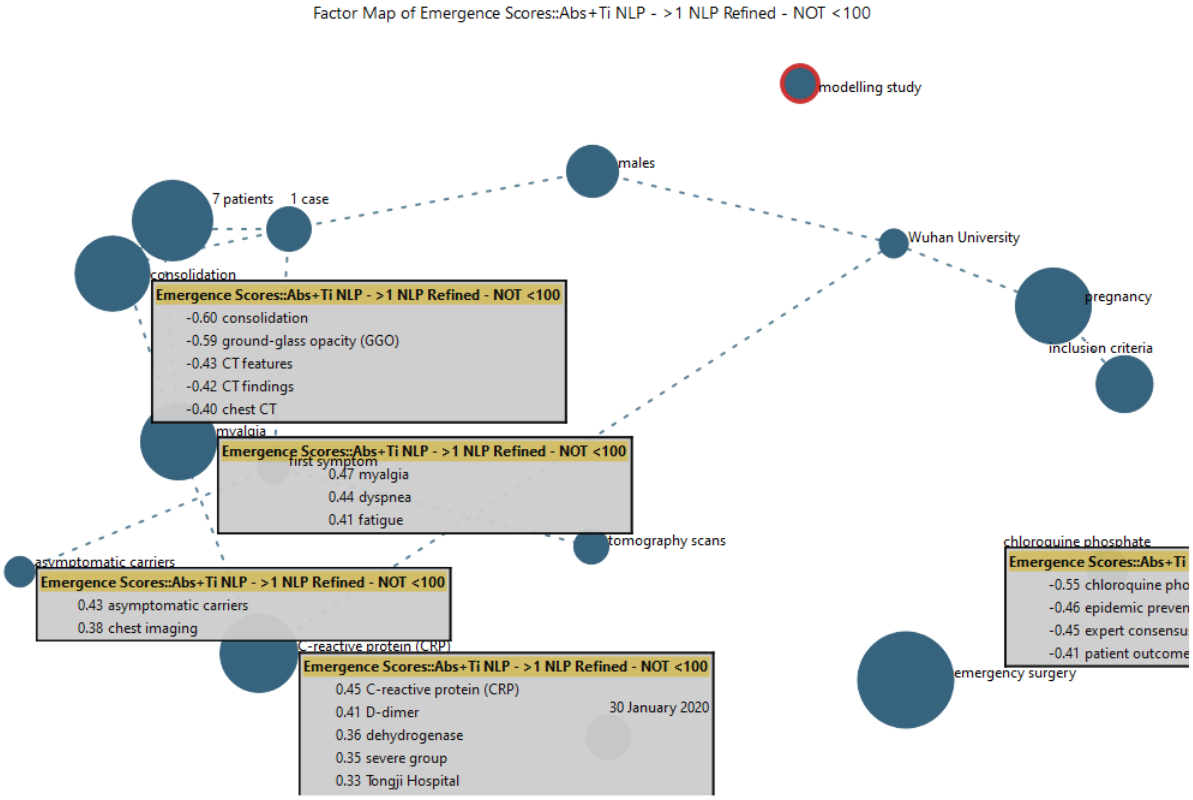
Exploring research interests; this dataset can help answer basic "**who, what, where, when?"** questions. Juxtaposing the available fields of data can point to research activity of particular concern:

- 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. Our profile based on the April 8 search results presented such a matrix. It allows one to scan topical frequency over time.
- 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).

- Figure 3 illustrates another perspective on the research. *VantagePoint* calculates degree of emergence of terms, spotlighting topics accelerating in recent attention. We extracted 343,166 abstract and title NLP (Natural Language Processing) terms and phrases from the records. We cleaned and consolidated those to 47,245 appearing in >1 and <100 records. Then we ran the software's Emergence Scoring routine that gave 465 emergent terms (showing high growth and meeting several criteria). In Figure 3 we group those using *VantagePoint's* factor mapping routine. Here we show the high-loading terms or phrases (ignore whether they are + or -) on several of the factors. We offer this to provoke thinking about possible ways you might probe this research corpus to identify cutting edge research.

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

Figure 3. Select Clusters of Emerging Topics



Background

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)"
- Also, among many resources: *NIH Extramural Nexus* listserv: https://list.nih.gov/cgi-bin/wa.exe?SUBED1=extramuralnexus&A=1