Hindsight2020: Characterizing Uncertainty in the COVID-19 Scientific Literature

Kinga Dobolyi[1]
corresponding author:kinga.dobolyi@gmail.com

George P. Sieniawski[2]
gsieniawski@iqt.org

David Dobolyi[3]
ddobolyi@nd.edu

Joseph Goldfrank[4]
goldfrank@stanford.edu

Zigfried Hampel-Arias[2]
zhampelarias@iqt.org

[1]formerly at IQTLabs, USA
[2]IQTLabs, USA
[3]University of Notre Dame, USA
[4]Stanford University, USA

July 23, 2021

Abstract

Background: In the healthiest of times, scientists face validity, methodology, and reproducibility challenges – following emerging infectious disease (EID) outbreaks, the rush to publish and the risk of data misrepresentation, misinterpretation, or misinformation puts an even greater onus on methodological rigor and proper understanding of the scientific method, which includes revisiting assumptions as circumstances change.
Methods: We apply claim-matching to various subsets of the COVID-19 scientific literature to identify synonymous scientific claims. Our goal is to build a framework for characterizing uncertainty in EID outbreaks as a function of time, peer review, hypothesis-sharing, evidence collection practice, and interdisciplinary citation networks.
Conclusions: This project seeks to understand how and when early evidence emerges for different types of recurring pathogen-related questions, outlined below, via a deep learning approach for claim-matching using SBERT. In addition, it makes publicly available an expert-annotated dataset of 5,815 matched sentence pairs that can be used to fine-tune future COVID-19 natural language programming models.

Keywords— SARS-Cov-2, uncertainty, NLP, research, policy, public health, pandemic

1 Introduction

While COVID-19 prompted a rapid surge in scientific research activity, several key questions remain unsettled over a year after the World Health Organization (WHO)'s pandemic declaration. For instance, although it is too early to characterize long-term disease sequelae, expected discoveries such as immunity duration and the risks of breakthrough infections also remain indeterminate [63]. Various transmissibility-related questions divided researchers for months in early 2020 [61], resulting in diminished trust in mask guidance among certain communities [67]. Since communicating uncertainty about emerging infectious disease outbreaks is inherently difficult, scientists and policy-makers facing outbreaks typically use a diverse set of approaches for distilling insights, acknowledging evidence gaps, updating public health guidance, and adjusting mitigation measures over time [11, 65].

Among these approaches is the Department of Homeland Security (DHS) Master Question List (MQL), discussed below, which outlines known unknowns about a wide variety of novel pathogens [1]. In addition to the MQL, related approaches like Grading of Recommendations Assessment, Development and Evaluation (GRADE)[62] can help global health organizations like WHO formulate outbreak response strategies *over a realistic range of time frames, while taking into account varying levels of evidence quality.*

Integrated assessment frameworks such as GRADE and the Master Question List (MQL) are particularly useful to tackle the information overload problem associated with infectious disease outbreaks: the deluge of academic articles and preprints published immediately after the onset stage. Implicit in both GRADE and the MQL is the idea that not all pandemic-related questions yield equally certain or equally prompt answers.

Some, like those involving randomized controlled trials of vaccines, take a while to gather sufficient data to resolve. Others, like those pertaining to decontamination, require relatively modest time investments to answer, both for the purposes of initial guidance and over longer term time-frames. While the public demands prompt answers to both types of questions, scientific researchers and health officials often face acute "exploitation-exploration" tradeoffs (i.e. acting on current knowledge despite uncertainty vs. investing in new discoveries to reduce uncertainty) [8]. To compound this challenge, vanity articles [34] can sometimes dominate headlines during the earliest stages of outbreaks, as was the case with COVID-19. In parallel, opinion pieces lacking novel results can also overshadow bona fide hypothesis development within the onset-stage pandemic literature [47]. While one might assume these publications are harmless noise, papers without novel results in fact can confound Natural Language Processing (NLP) pipelines which are often used to filter these massive article corpora, making information retrieval and sensemaking all the more challenging for human decision-makers.

1.1 Motivation

In this work, we examine the evolution of useful information on novel pathogens within scientific literature. This research seeks to understand how and when early evidence emerges for different types of recurring pandemic questions, such as those in the Department of Homeland Security Master Question List. Our goal is to build a framework for characterizing uncertainty in research on emerging infectious disease outbreaks as a function of time, impact, topic area, peer review, hypothesis-sharing, evidence col-

lection practice, and interdisciplinary citation networks, consistent with GRADE [62]. To do so, we trace human-curated evidence through scientific publications on SARS-CoV-2 over time to generate timelines around when questions are addressed, and how answers evolve. We begin by applying claim-matching, via NLP, to the roughly 600 sentences of evidence DHS cites in reply to the 16 MQL questions [1]. We seek to match each of these hundreds of sentences to similar and related claims in the COVID-19 Open Research Dataset (CORD-19) corpus [71] of 13 million sentences mined from the scientific literature on SARS-CoV-2. We then analyze the timing and uncertainty of new evidence over the course of the pandemic. This preliminary framework provides a foundation for global health experts and policy-makers during the onset phase of emerging pathogen outbreaks, as we aim to characterize when recurring questions about different types of diseases might begin to yield reliable answers.

In addition, we make public a dataset of 5,814 sentence pairs at our Hindsight2020 Github repository[1], each annotated with expert human judgements about the similarity utility between them, which we used for our modeling. This novel dataset expands upon previous Semantic Textual Similarity (STS) [29] and Question-Answering [45] work for COVID-19, by providing a unique and complementary annotation of sentence pairs requiring deeper biomedical knowledge than generally available through Amazon Mechanical Turk participants as in [29].

Specifically, we rank sentence similarity in terms of how closely-aligned two sentences are from the standpoint of research claim-matching: that is, if we wanted to find which papers contained sentences *similar in spirit* to the original evidence, we'd include statements describing the same phenomenon *even if they arrived at opposing research conclusions*. These ratings do not necessarily overlap with traditional STS scoring, in some sense, as many/most sentences would be rated *very similar* given that they all pertain to COVID-19. Similarly, our approach takes into account seemingly unrelated annotations in our dataset which may still be discussing the same topic or sub-topic (see rating scale in Table 1 along with examples in Section 3.5).

2 Related Work

2.1 Pandemic Uncertainty and Risk Communications

While the COVID-19 literature has grown exponentially since January 2020 [23], only 10% of preprints on COVID-19 resulted in publication in peer-reviewed journals, and machine learning methods are so-far unsuccessful at predicting which are likely to meet the standards of publication [6]. Furthermore, while select COVID-related articles may have appeared in the press earlier due to accelerated, and in some cases suspended, peer review, the public may have been presented results with a much higher risk for bias than what the same journals typically accept [21, 55, 74, 49]. In addition, between the start of the pandemic and May 2020, the majority of published research did not contain original data (for example, opinion pieces) [55]. This publication landscape presents acute risks for evidence-based public policy in the age of social media cf. [38].

Machine learning has emerged as an important means of assisting policy-makers and public health officials filter data [43]. While automated tools exist, such as screening preprints based on transparency and reproducibility criteria [73] or analyzing state-level policy announcements against labelled tweets [66], further work is needed. Interestingly, COVID-19 publication growth appears to have reached its apogee in May

[1] see csv in https://github.com/IQTLabs/hindsight2020/blob/main/hindsight/paper$_r esults$/

2020 and subsequently trends downward to Nov 2020 [35], potentially indicating diminished levels of vanity publishing and more conjectural articles after roughly half a year. Despite the volume of publication, high-quality COVID-19 research takes effort to identify; wasteful research (e.g. research which is inadequately produced and/or reported) will likely exceed the 85% pre-pandemic estimates for academic research in general [27, 14]. For example, [51] uses hydroxychloroquine as a case study to discuss how duplicative and low-quality research adversely impacted scientific discourse.

Several researchers have investigated integrated assessment frameworks, including GRADE [62], around uncertainty for COVID-19 policy-making, clinical evidence evaluation, and risk communication [11]. However uncomfortable for policy-makers, acknowledging the uncertainty associated with scientific evidence is a source of credibility and a means of retaining public trust [26]. The converse also appears to be true: overconfidence can backfire. For example, Pearce argues an estimate of the virus doubling rate in the UK was presented with unwarranted certainty and subsequently downplayed via role conflation between knowledge producers and knowledge users within a scientific advisory group, potentially delaying lockdowns and resulting in thousands of deaths [50]. Properly communicating uncertainty can help manage expectations and reduce public backlash when the infectious disease modeling community updates guidance based on new information [37]. Additionally, information overload about pandemics like COVID-19 (which often involves conflicting evidence) [44] is a significant factor in this context. Unfortunately, most of the public health recommendations for COVID-19 (such as masking, hand-washing, quarantine, and maintaining physical distance) relied on less recent research during the earliest phases of the pandemic [65] when policy experts had to extrapolate from prior experiences with other pathogens.

Finally, research agendas for COVID-19 have been developed for the Environmental Health sciences [24], and Master Question Lists (MQLs) have been developed in concert with the White House and others [4] as well as the Department of Homeland Security (DHS) [1]. The latter is based off a template of research questions identified during the Ebola virus outbreak of 2014 [16]. DHS provides a weekly update detailing evidence and answers to these questions for the COVID-19 pandemic. We use the citations within the DHS COVID-19 MQL as our ground truth in this work, as this biblographic artifact comprises a human-curated set of important results in the largely polluted [23] deluge of COVID-19 literature.

2.2 Biomedical Text Information Retrieval

Researchers often use Natural Language Programming (NLP) to search large biomedical text corpora for sentences related to a target keyword or scientific claim. However, NLP approaches to this class of problems must balance the complexity required for converting natural language text into meaningful and rich word embeddings that machine learning models can use, against the requirement to search and compare these embeddings with millions of potentially relevant sentence matches. For example, deep learning models such as BERT [19] and its progeny seem to be highly skilled in converting sentences to meaningful word embeddings, but could require thousands of compute hours to match our $\approx$600 claims against millions of sentences from academic articles. Fortunately, SBERT [57] is a recent advancement that uses twin BERT-Networks to generate word embeddings one can compare using cosine similarity. SBERT is as accurate as its predecessor, with the advantage of a two order-of-magnitude reduction in computation time [57].

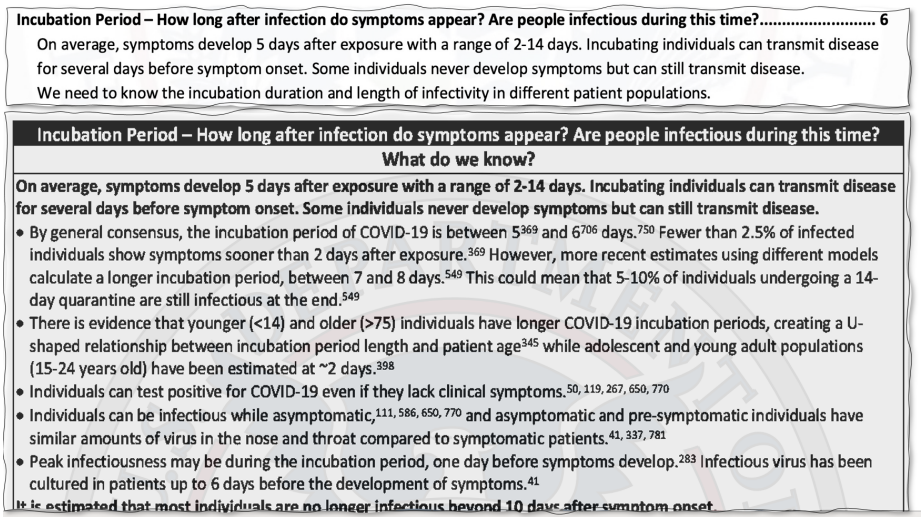
Various pre-trained deep learning models have been explored for text searching tasks, including *Semantic Textual Similarity (STS)*, *paraphrase identification*, *Question-Answering (QA)*, *claim-matching*. The models are typically pre-trained on large text corpora (such Bing searches and Wikipedia articles), and then can be fine-tuned on a domain-specific use-case via additional training on biomedical texts, for example. For this reason, various COVID-19 specific datasets have been developed to fine-tune these models for pandemic-related research filtering tasks. For instance, in [53], Pradeek et al. explore scientific claim verification on COVID-19 using VERT5ERINI. Their approach matches claims against article abstracts, which we ruled out for use in our work due to a high false negative rate we anticipated against our ground truth claims with this approach. In contrast, we do not focus solely on paper abstracts and we apply sentence similarity scoring to trace claims and their evolution through the COVID-19 literature.

In [48], Pal et al. built *EvidenceFlow* to predict emerging broad themes with respect to diseases (i.e. fever, diabetes, pneumonia) and biochemical substances (i.e. heparin, IL-6) within the COVID-19 literature using a novel approach for identifying critical nodes in entity networks weighted by cosine similarity [48, 3]. Unfortunately, their January 2021 predictions (respiratory failure, asthma, respiratory distress syndrome, acute respiratory syndrome coronavirus, tumor, acute kidney injury, dry cough, thrombotic, thromboembolic, lympohenia) lacked detail and specificity [3] to make these predictions meaningful and actionable for our purposes. Other claim-matching approaches have been applied to COVID-19 research which we also could have used for claim-matching in this research, but did not evaluate against our approach [70]. For example, [69] uses a Robustly Optimized BERT Pretraining Approach (RoBERTa) for claim-matching on CORD-19, while [25] also uses SBERT for related COVID-19 user queries.

2.3 COVID-19 NLP Datasets

As discussed above, pre-trained deep learning models are often fine-tuned on smaller, and typically more expensive, bespoke datasets, as domain-specific data require significant time and effort to gather. However, several annotated COVID-19 datasets have emerged in the past year specifically for this purpose. For example, non-medical experts have annotated, and in some cases reformatted, 11,000 research abstracts using Amazon Mechanical Turk (AMT) crowdsourcing [32]. COVID-QA is a question-answer dataset of about 2,000 question-answer pairs annotated by volunteer experts from COVID-19 academic articles [45]. Finally, the dataset we make publicly available in this work is most similar to CORD-19STS [29], a collection of almost 14,000 AMT-annotated sentence pairs to expose different STS levels. Our approach is similar to CORD-19STS, but we rely on expert annotation instead of AMT-based crowdsourcing.

Unannotated COVID-19 datasets are also available, and are generally used as the corpora for search tasks. COVID-19 has been analyzed using various social media and Google Trends datasets [41, 60, 33, 59, 28, 40, 72, 30, 75, 15, 46, 42, 10, 58, 64]. Several other options for COVID-19 text mining also exist for search, augmented reading, exploration, knowledge base construction, clinical diagnostic support, question answering, and summarization; [70] surveys these approaches. Many bibliometric analyses have been performed on COVID-19 academic articles [18, 39, 22, 36, 9, 20, 47], which share retrospective insights about the most common topics of study and article types (e.g., randomized controlled trials, observational studies, modeling/simulation, exploratory analyses, opinion pieces, peer-reviewed vs preprints). While these au-



Incubation Period – How long after infection do symptoms appear? Are people infectious during this time?..... 6

On average, symptoms develop 5 days after exposure with a range of 2-14 days. Incubating individuals can transmit disease for several days before symptom onset. Some individuals never develop symptoms but can still transmit disease.

We need to know the incubation duration and length of infectivity in different patient populations.

Incubation Period – How long after infection do symptoms appear? Are people infectious during this time?

What do we know?

On average, symptoms develop 5 days after exposure with a range of 2-14 days. Incubating individuals can transmit disease for several days before symptom onset. Some individuals never develop symptoms but can still transmit disease.

- By general consensus, the incubation period of COVID-19 is between 5[369] and 6[706] days.[750] Fewer than 2.5% of infected individuals show symptoms sooner than 2 days after exposure.[369] However, more recent estimates using different models calculate a longer incubation period, between 7 and 8 days.[549] This could mean that 5-10% of individuals undergoing a 14-day quarantine are still infectious at the end.[549]
- There is evidence that younger (<14) and older (>75) individuals have longer COVID-19 incubation periods, creating a U-shaped relationship between incubation period length and patient age[345] while adolescent and young adult populations (15-24 years old) have been estimated at ~2 days.[398]
- Individuals can test positive for COVID-19 even if they lack clinical symptoms.[50, 119, 267, 650, 770]
- Individuals can be infectious while asymptomatic,[111, 586, 650, 770] and asymptomatic and pre-symptomatic individuals have similar amounts of virus in the nose and throat compared to symptomatic patients.[41, 337, 781]
- Peak infectiousness may be during the incubation period, one day before symptoms develop.[283] Infectious virus has been cultured in patients up to 6 days before the development of symptoms.[41]

It is estimated that most individuals are no longer infectious beyond 10 days after symptom onset.

Figure 1: DHS Master Question List.

thors offer interesting insights, earlier bibliometric work generally does not factor in expert-identified utility and/or novelty.

3 Construction of HindSight2020 Dataset

3.1 Ground-truth claims from the DHS MQL

The DHS updates its Master Question List (MQL) citations on an ongoing basis, and we obtained an update from December 21, 2020, which provided expert-curated evidence to answer one of sixteen questions (see Figure 1). Almost 600 sentences of evidence were provided to answer these sixteen questions. For example, the claim *Individuals can be infectious while asymptomatic [111, 586, 650, 770], and asymptomatic and pre-symptomatic individuals have similar amounts of virus in the nose and throat compared to symptomatic patients [41, 337, 781]* is listed as one of around 40 sentences of evidence under the question *Incubation Period – How long after infection do symptoms appear? Are people infectious during this time?* These ≈600 sentences became our ground truth claims for this work, and each may be associated with one or more cited academic articles, trusted publications, news sources, and other materials.

3.2 Matching DHS Ground-truth claims against evidence in the CORD19 dataset of academic articles

We noted that only about half of the citations in the DHS MQL were directly from academic articles in CORD-19, the largest collection maintained of COVID-19 peer-reviewed articles and preprints we used to match against. CORD-19 is updated at least once a week by [71] at the Allen Institute for AI. Its articles are pre-processed into sentences and other metadata, available in *json* format. Thus, CORD-19 is frequently

and easily mined; for example, knowledge graph construction and knowledge discovery is an active area of research for COVID-19 [13, 7, 17, 54, 56, 68], with most of these on the CORD-19 dataset. We constructed our claim-matching dataset for SARS-CoV-2 academic sentence pairs by using a snapshot of the CORD-19 corpora we obtained on Jan 4, 2021. Sentences from both article abstracts and bodies were included for all articles and/or preprints that appeared in the ≈144,000 *pdf* article parses available. After filtering out articles that were older than 2020 (as the CORD-19 dataset includes articles on diseases potentially related to COVID-19, such as Ebolavirus and influenza), we obtained over 13 million sentences against which to compare our claims.

3.3 Filtering the CORD-19 dataset of academic articles

Next, we sought to filter these ≈13 million sentences from CORD-19 into a much smaller subset of potentially matching claims for each of the ≈600 ground truth sentences from the DHS MQL. To do so, we applied Named Entity Recognition (NER) using *spacy*'s [31] pre-trained *en core sci sm* model to identify relevant keywords in each of the DHS sentences. For each DHS sentence, we filtered the CORD-19 sentences to only those that had at least three uncased keyword matches between the two sentences; this further reduced our computation time to about a day to run our experiment. Twenty-nine sentences generated more than 80,000 filtered sentences to match against, which we set as a cutoff to preserve runtime; their keywords were manually reduced to have their pool of potential matches fall below the cutoff. Finally, 23 sentences (such as *Reinfection is possible.*) had too few keywords to match, so we relaxed the filter cutoff to include all sentences that matched two keywords for this small subset.

3.4 Mining matched claims using SBERT

Once each of the ≈600 DHS claims had its own set of up to 80,000 sentences to match against from the CORD-19 dataset, we used SBERT [57] to perform claim-matching. To do so, we experimented with a number of the different pre-trained models available [2], and finally settled on the *msmarco distilbert base v2* question-answer retrieval model over a STS model, as it seemed to generate better matches than the latter on a pilot study. We initially surmised STS would be a best-fit for our claim-matching needs, but as our goal was not necessarily to match exact claims in the literature (as the DHS sentences already typically came with citations for this purpose), we envisioned such a QA model (pre-trained on Bing searches) to allow for more interesting matches where other papers may cite or investigate similar topics to our ground truth evidence. We configured SBERT to provide up to the top ten matches for each of our ≈600 ground truth claims.

In our results, we often were able to trace back non-academic DHS citations to older academic sources. In cases where citations were available from DHS that existed in our CORD-19 database of academic articles, our claim-matching approach (described in this section) was able to directly match the cited article about 20% of the time; for the remaining 80%, our expert annotators were able to manually evaluate similarity.

While such a low initial match might seem like a disappointment (and probably is for claim-matching researchers), in our case it was actually desirable, as our goal was not to prove that we could pinpoint the original papers (as we already had access to DHS citations), but that we could trace the evolution of those claims through subsequent research, either with similar research results in another paper, or a mention

Table 1: Annotator Rating Scale

Rating	Interpretation
Yes, definitely	Both sentences describe the same phenomenon
Perhaps	The second sentence is conceptually 'related work'
No	Although they may share the same topic, the two sentences are not describing the same things

in a related work section. Often, DHS-cited claims were paraphrased or summarized to the point of an inability for us to detect any meaningfully related sentences (such as *Reinfection is possible.*); this occurred in 4% of the DHS sentences. 346 of the ≈600 ground truth claims had at least one paper cited directly by DHS found in the CORD-19 dataset (though not necessarily matched by our algorithm). By and large, the remaining DHS citations were non-academic articles (e.g. news sources like Reuters or *The New York Times*).

3.5 Expert annotation of HindSight2020 matched claim pairs

Next, we evaluated the quality of the matched claim-sentence pairs using human experts, to ensure that further analysis was meaningful. While SBERT [57] is a powerful, automated tool for claim-matching and related tasks, it often generates false-positives. As the goal of our claim-matching is to trace the evolution of evidence through academic literature, models fine-tuned for more traditional STS and QA tasks can fail to appreciate the subtlety required for our similarity ranking. For example, distance from a topic is too broad to quantify similarity in our approach, as most sentences are potentially *related* as they discuss SARS-CoV-2/COVID-19. Instead, we relied on expert human annotators to perform the following scoring in 1. 2 shows examples of the rating scale applied to matched sentence pairs.

Sentence-scoring is an extremely resource-intensive, expensive, and arguably subjective process, even for experts (much like the categorization decisions journal editors and MQL compilers make). We relied on a single expert annotator to rate all 5,814 sentences initially; this bioinformatician had extensive background in the COVID-19 literature, following the academic articles closely for the past year. Since our goal was not to evaluate the quality of the research, but rather, to judge whether two sentence described the same phenomenon, it does not require deep biological expertise. Table 2 is an example of the set of ten potential matches for the DHS claim *This could mean that 5-10% of individuals undergoing a 14-day quarantine are still infectious at the end*, along with their expert-judged research-similarity rating.

We suggest researchers consider our annotations as more of a continuous range, rather a strict binning. Finally, the DHS evidence for forecasting models (DHS Question 16), yielded low-quality matches via SBERT, possibly because the DHS language here was a bullet list of incomplete sentences, which could be more difficult to match for similarity. The timeline for this question may not fully incorporate early, unpublished, attempts at modeling the progression of the pandemic.

Table 2: Annotator Rating Scale examples

DHS sentence	This could mean that 5-10% of individuals undergoing a 14- day quarantine are still infectious at the end.
Rating	**CORD19 sentence match to DHS sentence above**
Yes, definitely	Notably, even a 14-day quarantine period does not eliminate the risk of individuals spending time infectious after release.
Perhaps	Under a 5-day quarantine period, around 6.8% of infected arrivals are released while highly infectious.
No	However if we assume anyone who is infected by the virus will be under quarantine 7 days after s/he becomes infectious, then our model estimate for R0 2.8-3.6 which is comparable to the current popular estimates of R0 reported in the literature.

4 Experimental Results

In order to analyze the timing patterns of evidence presentation and collection for the DHS MQL, we filtered the annotated sentence pairs above to include only the set of sentences our annotators labelled as *yes, definitely* or *perhaps* (sentences describing the same phenomenon or conceptually related work). We also manually filtered out sentences which appeared to be duplicative citations or references to earlier studies, as we wanted to minimize the impact of these types of sentences. The analysis below presents our findings on this subset of evidence (i.e. close matches of original research against the original DHS claims).

4.1 Hypothesis 1: Different questions are answered at different times

Our first hypothesis was that high-quality evidence in CORD-19 for the sixteen questions from the DHS MQL would emerge at different times, and not match the pattern of an exponential growth in publication from January through May 2020, followed by a slow-but-steady decline through the rest of that calendar year. We graphed all of the statements filtered as described above, shown in Figure 2.

While we expected evidence involving vaccines and protective immunity to accumulate later in time, we were surprised to see that even questions around PPE, transmissibility, clinical diagnosis, and environmental stability continued to engage scientists for months, peaking in the second half of 2020. While it remains possible that we failed to filter out statements from our evidence that referred to earlier research (as the natural language in academic articles may not always explicitly refer to another study or include a citation), we also know in hindsight that the research community revised many of the answers to these questions in light of subsequent findings, as we discuss in section 5.1.

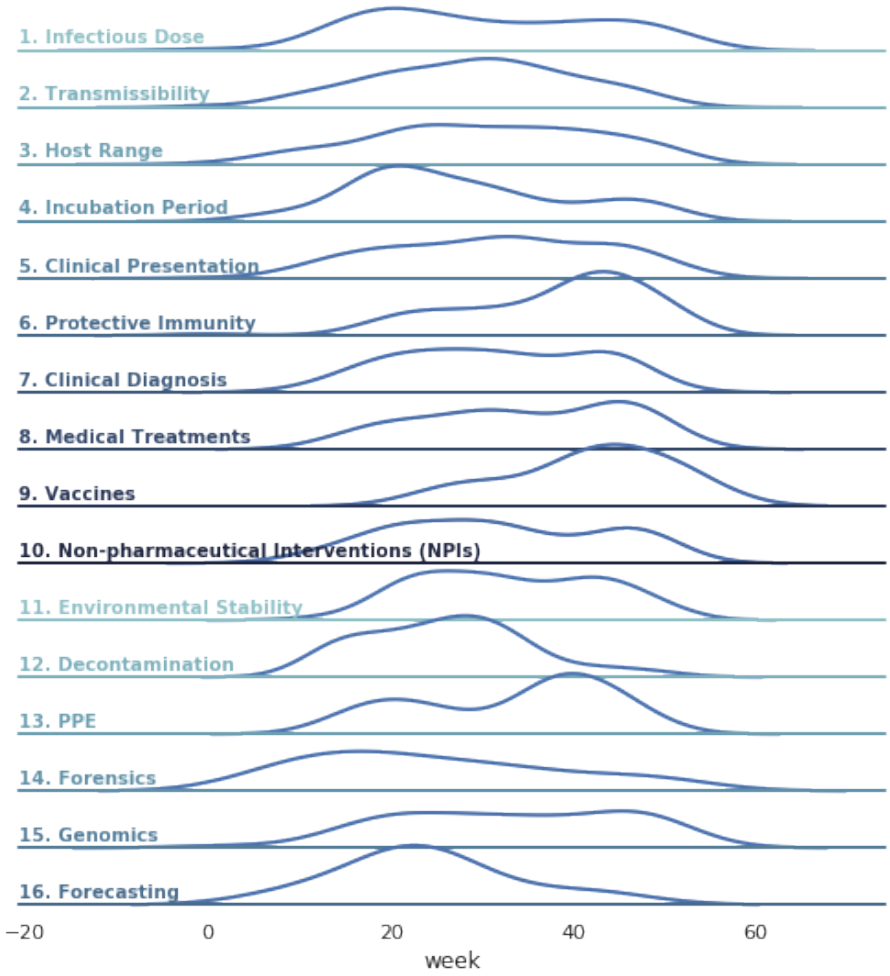


Figure 2: Timing patterns of CORD-19 matches of original research

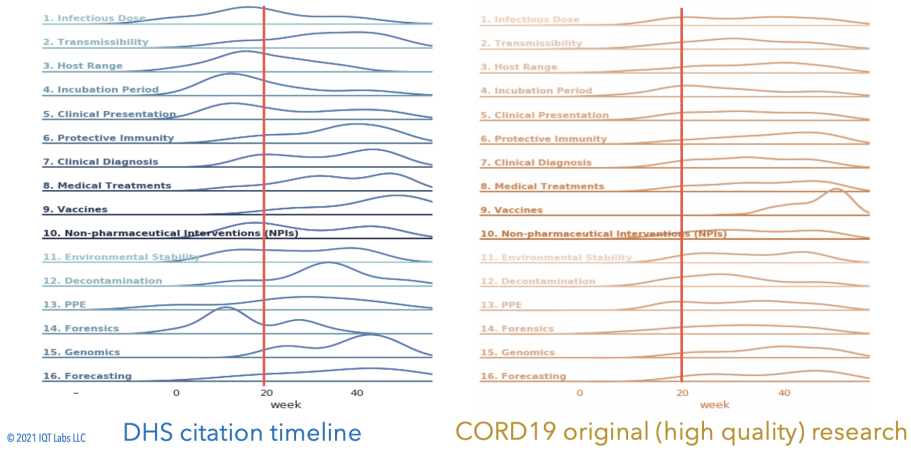


Figure 3: Timing patterns of DHS evidence vs CORD-19 close matches of original research

We next compared the timing of DHS claims against the timing of close matching evidence (as filtered above, but with the additional restriction of including only *yes* labels) in the CORD-19 dataset. We hypothesized that the DHS claims would occur earlier, on average, than when these questions would be answered in CORD-19. Indeed, every question in the MQL seems to be answered no earlier in CORD-19 than by DHS, with the exception of *Decontamination*, as shown in Figure 3.

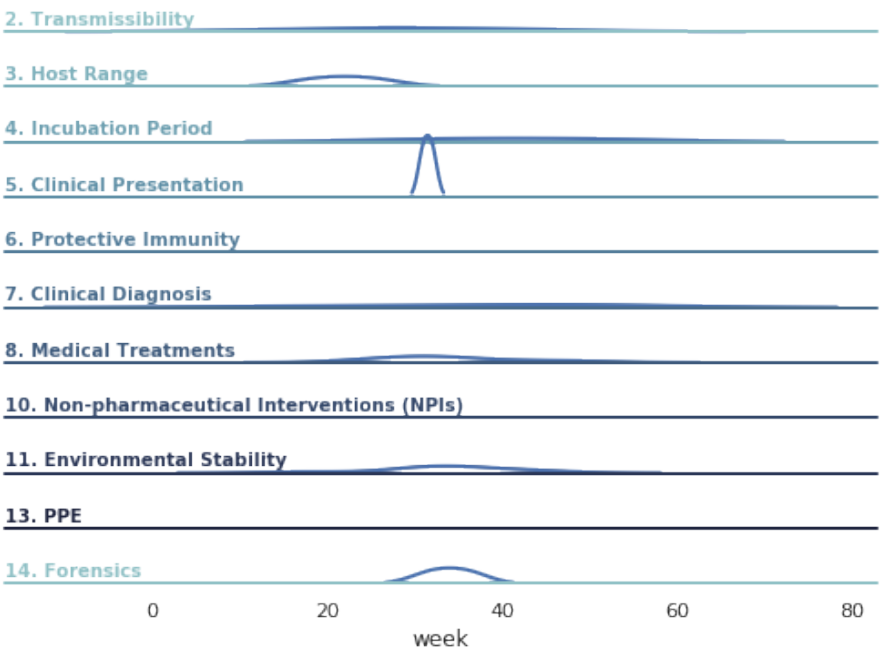
Overall, obtaining answers to the DHS MQLs as early as possible is the goal of effective public health policy-making, as it enables timelier crisis response and resource allocation decisions, ultimately saving lives and minimizing the impact of emerging disease outbreaks [11]. For SARS-CoV-2, it appears the the DHS MQL compilers were able to identify meaningful answers to these questions early and often, across a range of pandemic-related issues.

4.2 Hypothesis 2: Contradictory evidence appears after initial claims

Early MQL answers that do not later change are especially valuable. To measure how often this occurred, we performed an entailment vs contradiction analysis on our DHS-CORD-19 sentence pairs that were labelled as *yes* or *maybe* by our expert annotator. Using the MedNLIs glove-bio-asq-mimic[2] model, we predicted whether the matched claim is an entailment or contradiction of the DHS claim. Such an NLP approach generated many false positives, so we manually reviewed each pair labelled as an apparent contradiction, arriving at ≈40 actual contradictions in our dataset of 5,814 sentence pairs. We then graphed this subset, shown in Figure 4.

Of these ≈40 contradictions that had the paper cited by DHS in our CORD-19 dataset (and we could therefore note its publication date), 8 were cases where the DHS evidence cited the *new* research; very rarely did the original DHS conclusions change in the future. Most contradictions could be found within ≈10 to ≈30 weeks from the

[2]https://github.com/jgc128/mednli_baseline



c

Figure 4: Timing patterns of results contradicting DHS claims.

start of the pandemic, and overall they were rare. A closer inspection of the sentence pairs that revealed contradictions included the specific topics of aerosol transmission, pre- and/or a- symptomatic transmission, the infectiousness of children, the benefits of certain investigational drugs (such as anankira, favipiravir, and hydroxychloroquine), environmental surface contamination, and pangolins as intermediate hosts. While one would expect, over time, for the value of repurposing various drugs to change, as case studies and smaller cohorts might progress to randomized clinical trials, it is less clear why uncertainty related aerosol transmission and pre-symptomatic spread was not recognized sooner; perhaps the presenting similarity of SARS2 to a more traditional respiratory illness like influenza, or the biological similarity to SARS1, biased researchers at the start of the pandemic.

While most evidence did not change over time, there are some potential lessons learned for future pandemics. The evidence for mechanisms of transmission and clinical presentation of emerging diseases may change over the first few months of a pandemic, and recommending low-cost nonpharmaceutical interventions (NPIs) such as masking and voluntary social distancing early and consistently could help mitigate the impact from future respiratory disease outbreaks, until the consensus around these question areas stabilizes. Similarly, initial success of repurposed medications should be taken with a grain of salt, as larger subsequent clinical trials may reveal an absence of benefit. Overall, however, answering the DHS questions early seems like a feasible and highly useful approach, especially if one is able to characterize the certainty around how those answers might change, as we are proposing to do later.

4.3 Hypothesis 3: Evolution in certainty of claims varies across questions

Our results show that the research published early during a pandemic, despite all of its limitations, can be successfully curated by a human into a set of early, actionable evidence, with the caveat that some MQL answers are more likely to change than others. Next, we wanted to investigate the language scientists use within their publications to communicate this uncertainty: does it change over time? Do scientists write with more conviction early on in response certain questions? Alternatively, do they hedge using verbal quantifiers (e.g., "potentially", "somewhat likely," etc.), expressing hesitancy about initial results pending subsequent validation? Knowing the breakdown of certainty in the language of evidence across questions and over time is helpful for humans (and machines) curating these claims, as it might prevent unintentional bias in the interpretation of the confidence of early research results.

To investigate the evolution of certainty expressed in the evidence from academic articles over time, we used the Linguistic Uncertainty Classifier Interface by Vincze et al. [5] to label the language of every DHS and CORD-19 sentence of evidence as either *certain* or *uncertain*. We then graphed each DHS-CORD-19 sentence pair over time, using the CORD-19 paper date, to show how the DHS evidence to CORD-19 evidence transition over time: either *certain to certain*, *certain to uncertain*, *uncertain to certain*, or *uncertain to uncertain*. We only graph sentence pairs below where the CORD-19 evidence came later than the DHS evidence (when we had a date available for the latter).

Each tile in Figure 5 represents the evolution of uncertainty over time, per question, between the earlier DHS claim and the matching CORD-19 evidence. Points with the value U-U on the y-axis represent stability in uncertainty between the sentence pair;

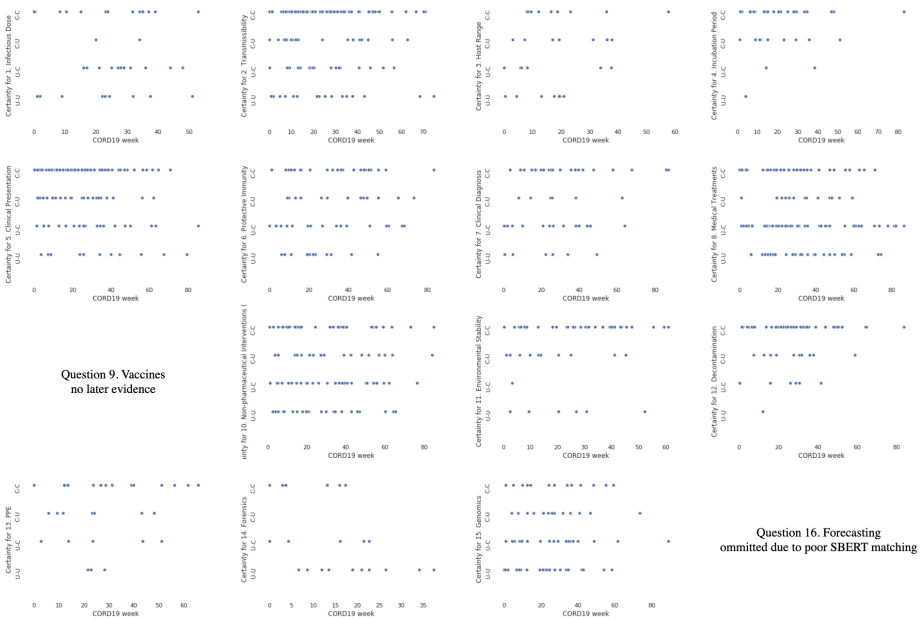


Figure 5: Uncertainty evolution over time between DHS claims and matching CORD-19 evidence from a later date.

U-C implies a transition growing from uncertain to certain; C-U is a reduction in certainty over time; and C-C represents stability in certainty between pairs. We had no points to graph for Question 9 (Vaccines) and Question 16 (Forecasting) due to a lack of matching CORD-19 papers or evidence with timestamps from DHS.

Overall, many questions seemed to express certainty throughout the pandemic more often that not, including incubation period, clinical presentation, environmental stability, and decontamination. Meanwhile, transmissibility, medical treatments, NPIs, and genomics had more frequent uncertainty in the language of their claims in terms of total number of *uncertain to uncertain* sentence pairs. In terms of changes, medical treatments had the most instances of moving from uncertain to certain language between sentence pairs. Finally, the evolution we were most concerned about, a change from *certain to uncertain* (represented by the second-from-top row in each graph) was less frequent in general than the three other types of potential evolution. For reference, other studies have shown that most papers that evolved from preprints to journal publications were largely similar in reporting of study characteristics, outcomes and spin [12].

5 Discussion

Given the deluge in academic preprints and peer-reviewed papers being published on SARS-CoV-2 last year [23], we sought to determine if it was possible to extract reliable answers to outbreak-related questions from this early literature. While topics such as vaccines require months (if not years) to mature into usable research outcomes, we

were curious what happened to early evidence mined to answer other types of questions (e.g., about clinical presentation, transmission, and decontamination).

Overall, we found that most early human-curated evidence DHS compiled into the Master Question List (MQL) was highly reliable and stable over time. We also were able to document that when newer evidence contradicted original conclusions, this generally happened within 2 to 6 months of the start of the pandemic. Therefore, it seems that the highest risk of evidence changing occurs in the first six months of a novel outbreak, and policy-making could aim to be more conservative at first (for example, assuming masks are needed), while preparing to relax restrictions and recommendations after the six-month window has passed. Some academic texts on the same topic moved from certain to uncertain language, but this was rarer than other types of certainty language evolution (or stability). In building our framework of uncertainty, we found this shift correlated with contradictions in the literature.

Our analysis can help provide timelines for public health officials navigating the challenges around incomplete information (in other words, situations in which the *absence of evidence is not evidence of absence*). Having an understanding of MQL citation timelines for specific topics can provide estimates for when we can reliably know enough stable information to identify and implement stable policy; before such a time, more conservative measures can be installed with the explicit caveat that they will be reviewed and eased as more information becomes available *during a more or less predictable timeline.*

Taken together, we have laid the foundations for building a framework around uncertainty in research evidence from academic articles on novel pathogens by 1) establishing that different themes and questions are answered at different times, which do not always correlate with paper publication volumes; 2) while the likelihood of change for early conclusions is low overall, we are able to bound these changes to be between 2 and 6 months from the start of the pandemic; 3) we can quantify which question topics tend to evolve to be more, or less, certain over time in terms of the language researchers use to describe their conclusions. Our next steps would be to determine the impact of a potential piece of evidence being incorrect, given how long since the start of the pandemic, and how many people have been infected and impacted so far; we plan to address these elements in future work.

5.1 Limitations

There are several limitations to the work presented here, including limitations of claim-matching. For example, we may have missed matching claims in CORD-19 due to the DHS ground truth sentence being a paraphrase, or SBERT and/or our NER tools potentially not recognizing – e.g. a *migraine* might be equivalent for our purposes to a *headache*, e.g. These and similar limitations open the possibility of false negatives in our automated claim-matching approaches. Given the subjectivity of our expert annotations, it is also possible that there are additional false positive and false negative matches. Another limitation was our reliance on a single expert annotator to decide the quality of SBERT matches, due to the expense of this task. Although our preliminary results indicate that when the same expert re-rates samples of the ten matches for ten arbitrary DHS claims (100 sentence-pairs total), they arrive at the same rating 95% of the time, while a separate annotator we had on hand agreed with our expert 85% of the time, we seek to formally measure intra- and inter-annotator agreement in an ongoing followup to this work.

In addition, the DHS MQL is a living document, and we only analyzed its ground

truths for a single snapshot; it is possible that the reason we found its evidence to be so stable is that it already had been revised over time. However, during our certainty and contradiction analyses, we only examined relationships where the publication date was available for the ground truth sentence to us, to try to obviate this concern. It is still possible that select evidence which turned out to be wrong was removed; we propose to explore such evolution of weekly updates to the MQL in future work.

Finally, we only examined the timelines for evidence collection across DHS questions for SARS-CoV-2 in this work. The timelines for other pathogens and outbreak scenarios may differ substantially. To study whether our approach generalizes, we propose to repeat our experiments on the Ebolavirus literature, using DHS' MQL for the 2014 Ebola outbreak in West Africa. In [52], Porter et al. have noted a low semantic similarity between research topics for SARS-CoV-2 generated in 2020 and earlier research, which might indicate novel and/or unique topics for this outbreak (and possibly future ones) that may not fit neatly into existing MQL frameworks.

5.2 Conclusion

In this work, we created the foundations for a framework to characterize uncertainty around evidence in the context of academic articles on emerging infectious diseases. Our goal was to understand how different questions, such as those around transmissibility versus vaccines, for example, have evidence that changes and occasionally contradicts itself over time, and to be able to predict when and where these reversals may occur within the scientific literature. This framework can help inform policy at the onset and post-peak stages of infectious disease outbreaks, as it can quantify, both in time and in impact, when existing evidence may be likely to change, and accordingly, where carefully crafted public health risk communications may be most critical.

5.3 Acknowledgements

We thank Dylan George, Dan Hanfling, Kevin O'Connell, Benjamin Lee, and Nina Lopatina for their feedback and suggestions on this research, and Ben Rocklin for his SBERT-based claim-matching code that we adopted for our experiments. Nina Lopatina was also a contributor to the claim-matching repository as the project lead.

References

[1] URL `https://www.dhs.gov/publication/st-master-question-list-COVID-19`.

[2] URL `https://www.sbert.net/docs/pretrained_models.html`.

[3] URL `https://evidenceflow.tavlab.iiitd.edu.in/emerging.html`. examined on 2/11/2021.

[4] URL `https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks`.

[5] URL `https://github.com/meyersbs/uncertainty/wiki/installation-&-usage`.

[6] A. Älgå, O. Eriksson, and M. Nordberg. The Development of Preprints During the COVID-19 Pandemic. *J. Internal Med.* doi: doi.org/10.1111/joim.13240.

[7] A. Amini, T. Hope, D. Wadden, M. van Zuylen, E. Horvitz, R. Schwartz, and H. Hajishirzi. Extracting a Knowledge Base of Mechanisms from COVID-19 Papers, 2020. URL `https://arxiv.org/abs/2010.03824`.

[8] D. Angus. Optimizing the Trade-off Between Learning & Doing in a Pandemic. *JAMA*, May 2020. doi: doi:10.1001/jama.2020.4984.

[9] S. Aviv-Reuven and A. Rosenfeld. Publication Pattern Changes Due to the COVID-19 Pandemic: a Longitudinal & Short-term Scientometric Analysis. 2021. URL `https://arxiv.org/abs/2010.02594`.

[10] A. Bansal, S. Gupta, V. Jain, A. Kumar, and A. Klein. Utilizing Google Trends to Assess Worldwide Interest in COVID-19 & Myocarditis. *J. Med. Sys.*, 45(2), Jan. 2021. doi: 10.1007/s10916-020-01685-5.

[11] L. Berger, N. Berger, V. Bosetti, I. Gilboa, L. P. Hansen, C. Jarvis, M. Marinacci, and R. D. Smith. Rational Policymaking During a Pandemic. *Proc. of the Nat'l Acad. of Sci.*, 118(4), 2021. doi: 10.1073/pnas.2012704118.

[12] L. Bero, R. Lawrence, L. Leslie, K. Chiu, S. McDonald, M. J Page, Q. Grundy, L. Parker, S. L Boughton, J. J Kirkham, and R. Featherstone. Comparison of Preprints & Final Journal Publications From Covid-19 Studies: Discrepancies in Results Reporting & Spin in Interpretation. *medRxiv*, 2021. doi: 10.1101/2021.04.12.21255329.

[13] G. Cernile, T. Heritage, N. J. Sebire, B. Gordon, T. Schwering, S. Kazemlou, and Y. Borecki. Network Graph Representation of COVID-19 Scientific Publications to Aid Knowledge Discovery. *BMJ Health & Care Informatics*, 28(1):e100254, Jan. 2021. doi: 10.1136/bmjhci-2020-100254.

[14] I. Chalmers and P. Glasziou. Avoidable Waste in the Production & Reporting of Research Evidence. *The Lancet*, 374(9683):86–89, Jul. 2009. doi: 10.1016/s0140-6736(09)60329-9.

[15] R. Chandrasekaran, V. Mehta, T. Valkunde, and E. Moustakas. Topics, Trends, & Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study. *J. Med. Internet Res.*, 22(10):e22624, Oct. 2020. doi: 10.2196/22624.

[16] L. A. Colf, R. Brothers, and C. E. Murata. A Role for Science in Responding to Health Crises, Aug. 2016.

[17] J. Collard, T. Bhat, E. Subrahmanian, I. Monarch, J. Tash, R. Sriram, and J. Elliot. A Web Resource for Exploring the CORD-19 Dataset Using Root- & Rule-Based Phrases. *J. the Indian Inst. of Science*, 100(4):725–731, Sept. 2020. doi: 10.1007/s41745-020-00193-2.

[18] H. Dehghanbanadaki, F. Seif, Y. Vahidi, F. Razi, E. Hashemi, M. Khoshmirsafa, and H. Aazami. Bibliometric Analysis of Global Scientific Research on Coronavirus. *Med J.*, May 2020. doi: doi:10.34171/mjiri.34.51.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of the 2019 Conf. of the North American Chapter of the Ass'n for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long & Short Papers), pages 4171–4186, Minneapolis, Minnesota, Jun. 2019. Ass'n for Computational Linguistics. doi: 10.18653/v1/N19-1423.

[20] A. Doanvo, X. Qian, D. Ramjee, H. Piontkivska, A. Desai, and M. Majumder. Machine Learning Maps Research Needs in COVID-19 Literature. *Patterns*, 1(9): 100123, Dec. 2020. doi: 10.1016/j.patter.2020.100123.

[21] I. Y. Elgendy, N. Nimri, A. F. Barakat, J. Ibrahim, J. Mandrola, and A. Foy. A Systematic Bias Assessment of Top-cited Full-length Original Clinical Investigations Related to COVID-19. *Eur. J. Internal Med.*, 2021. doi: doi: 10.1016/j.ejim.2021.01.018.

[22] H. ElHawary, A. Salimi, N. Diab, and L. Smith. Bibliometric Analysis of Early COVID-19 Research: The Top 50 Cited Papers. *Infectious Diseases: Res. & Treatment*, 13:1178633720962935, 2020. doi: 10.1177/1178633720962935.

[23] H. Else. How a Torrent of COVID Science Changed Research Publishing —in Seven Charts. *Nature*, 588(7839):553–553, Dec. 2020. doi: 10.1038/d41586-020-03564-y.

[24] N. A. Errett, M. Howarth, K. Shoaf, M. Couture, S. Ramsey, R. Rosselli, S. Webb, A. Bennett, and A. Miller. Developing an Environmental Health Sciences COVID-19 Research Agenda: Results from the NIEHS Disaster Research Response (DR2) Work Group's Modified Delphi Method. *International J. Env. Res. & Pub. Health*, 17(18):6842, Sept. 2020. doi: 10.3390/ijerph17186842.

[25] A. Esteva, A. Kale, R. Paulus, K. Hashimoto, W. Yin, D. Radev, and R. Socher. CO-Search: COVID-19 Information Retrieval with Semantic Search, Question Answering, & Abstractive Summarization. 2020. URL `https://arxiv.org/abs/2006.09595`.

[26] S. T. Fiske and C. Dupree. Gaining Trust as Well as Respect in Communicating to Motivated Audiences About Science Topics. *Proc. Natl. Acad. Sci.*, Sept. 2014. doi: doi:10.1073/pnas.1317505111.

[27] P. P. Glasziou, S. Sanders, and T. Hoffmann. Waste in COVID-19 Research. *BMJ*, 369, 2020. doi: 10.1136/bmj.m1847.

[28] J.-W. Guo, C. L. Radloff, S. E. Wawrzynski, and K. G. Cloyes. Mining Twitter to Explore the Emergence of COVID-19 Symptoms. *Pub. Health Nursing*, 37(6): 934–940, Sept. 2020. doi: 10.1111/phn.12809.

[29] X. Guo, H. Mirzaalian, E. Sabir, A. Jaiswal, and W. Abd-Almageed. CORD19STS: COVID-19 Semantic Textual Similarity Dataset, 2020. URL `https://europepmc.org/article/PPR/PPR272247`.

[30] T. Hoang and P. Vu. Not-NUTs at WNUT-2020 Task 2: A BERT-based System in Identifying Informative COVID-19 English Tweets. pages 466–470, Nov. 2020.

[31] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL `https://doi.org/10.5281/zenodo.1212303`.

[32] T.-H. K. Huang, C.-Y. Huang, C.-K. C. Ding, Y.-C. Hsu, and C. L. Giles. CODA-19: Using a Non-Expert Crowd to Annotate Research Aspects on 10,000+ Abstracts in the COVID-19 Open Research Dataset. 1, Jul. 2020.

[33] T. Huynh, L. Thanh Luan, and S. T. Luu. BANANA at WNUT-2020 Task 2: Identifying COVID-19 Information on Twitter by Combining Deep Learning & Transfer Learning Models. In *Proc. of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 366–370, Online, Nov. 2020. Ass'n for Computational Linguistics. doi: 10.18653/v1/2020.wnut-1.50.

[34] R. Jalali, A. Hosseinian-Far, and M. Mohammadi. Contradictions in the Promotion of Publishing Academic & Scientific Journal Articles, & the Inability to Cope With the New Coronavirus (COVID-19). *Antimicrobial Resistance & Infection Control*, 10(1), Jan. 2021. doi: 10.1186/s13756-021-00884-0.

[35] M. Kang, S. S. Gurbani, and J. A. Kempker. The Published Scientific Literature on COVID-19: An Analysis of PubMed Abstracts. *J. Med. Sys.*, 45(1), Nov. 2020. doi: 10.1007/s10916-020-01678-4.

[36] K. Kousha and M. Thelwall. COVID-19 Publications: Database Coverage, Citations, Readers, Tweets, News, Facebook Walls, Reddit Posts. *Quantitative Science Studies*, 1(3):1068–1091, 2020. doi: 10.1162/qss_a_00066.

[37] S. E. Kreps and D. L. Kriner. Model Uncertainty, Political Contestation, & Public Trust in Science: Evidence From the COVID-19 Pandemic. *Science Advances*, 6 (43), 2020. doi: 10.1126/sciadv.abd4563.

[38] H. J. Larson. The Biggest Pandemic Risk? Viral Misinformation. *Nature*, Oct. 2018. doi: doi:10.1038/d41586-018-07034-4.

[39] J. Lever and R. B. Altman. Analyzing the Vast Coronavirus Literature With Coronacentral. *bioRxiv*, 2020. doi: 10.1101/2020.12.21.423860.

[40] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh. Natural Language Processing Reveals Vulnerable Mental Health Support Groups & Heightened Health Anxiety on Reddit During COVID-19: Observational Study. *J. Med. Internet Res.*, 22(10):e22635, Oct. 2020. doi: 10.2196/22635.

[41] T. Lu and B. Y. Reis. Internet Search Patterns Reveal Clinical Course of COVID-19 Disease Progression & Pandemic Spread Across 32 Countries. *npj Dig. Med.*, 4(1), Feb. 2021. doi: 10.1038/s41746-021-00396-6.

[42] J. C. Lyu and G. K. Luli. Understanding the Public Discussion About the Centers for Disease Control & Prevention During the COVID-19 Pandemic Using Twitter Data: Text Mining Analysis Study. *J. Med. Internet Res.*, 23(2):e25108, Feb. 2021. doi: 10.2196/25108.

[43] M. Mehraeen, M. Dadkhah, and A. Mehraeen. Investigating the Capabilities of Information Technologies to Support Policymaking in COVID-19 Crisis Management; A Systematic Review and Expert opinions. *Eur. J. Clin. investigation*, 50: e13391, 09 2020. doi: 10.1111/eci.13391.

[44] M. Mohammed, A. Sha'aban, A. I. Jatau, I. Yunusa, A. M. Isa, A. S. Wada, K. Obamiro, H. Zainal, and B. Ibrahim. Assessment of COVID-19 Information Overload Among the General Public. *J. Racial & Ethnic Health Disparities*, Jan. 2021. doi: 10.1007/s40615-020-00942-0.

[45] T. Möller, A. Reina, R. Jayakumar, and M. Pietsch. COVID-QA: A Question Answering Dataset for COVID-19. In *Proc. of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, Jul. 2020. Ass'n for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.nlpCOVID19-acl.18`.

[46] C. Murray, L. Mitchell, J. Tuke, and M. Mackay. Symptom Extraction From the Narratives of Personal Experiences With COVID-19 on Reddit. 2020. URL `https://arxiv.org/abs/2005.10454`.

[47] A. Odone, S. Galea, D. Stuckler, C. Signorelli, A. Amerio, L. Bellini, D. Bucci, M. Capraro, G. Gaetti, S. Salvati, A. Amerio, L. Bellini, D. Bucci, M. Capraro, G. Gaetti, and S. Salvati. The First 10 000 COVID-19 Papers in Perspective: Are We Publishing What We Should Be Publishing? *Eur. J. Pub. Health*, 30(5): 849–850, Aug. 2020. doi: 10.1093/eurpub/ckaa170.

[48] R. Pal, H. Chopra, R. Awasthi, H. Bandhey, A. Nagori, A. Gulati, P. Kumaraguru, and T. Sethi. Predicting Emerging Themes in Rapidly Expanding COVID-19 Literature with Dynamic Word Embedding Networks & Machine Learning. *medRxiv*, 2021. doi: 10.1101/2021.01.14.21249855.

[49] A. Palayew, O. Norgaard, K. Safreed-Harmon, T. H. Andersen, L. N. Rasmussen, and J. V. Lazarus. Pandemic Publishing Poses a New COVID-19 Challenge. *Nat. Hum. Behav.*, 2020. doi: doi:10.1038/s41562-020-0911-0.

[50] W. Pearce. Trouble in the Trough: How Uncertainties Were Downplayed in the UK's Science Advice on COVID-19. *Humanit. Soc. Sci. Comms.*, Oct. 2020. doi: doi.org/10.1057/s41599-020-00612-w.

[51] L. Perillat and B. S. Baigrie. COVID -19 & the Generation of Novel Scientific Knowledge: Research Questions & Study Designs. *J. Evaluation in Clin. Practice*, 27(3):694–707, Feb. 2021. doi: 10.1111/jep.13550.

[52] A. L. Porter, Y. Zhang, Y. Huang, and M. Wu. Tracking & Mining the COVID-19 Research Literature. *Frontiers in Res. Metrics & Analytics*, 2020. doi: 10.3389/frma.2020.594060.

[53] R. Pradeep, X. Ma, R. Nogueira, and J. Lin. Scientific Claim Verification with VERT5ERINI, 2020. URL `https://arxiv.org/abs/2010.11930`.

[54] P. Radanliev, D. D. Roure, and R. Walton. Data Mining & Analysis of Scientific Research Data Records on COVID-19 Mortality, Immunity, & Vaccine Development - in the First Wave of the COVID-19 Pandemic. *Diabetes & Metabolic Syndrome: Clin. Res. & Revs.*, 14(5):1121–1132, Sept. 2020. doi: 10.1016/j.dsx.2020.06.063.

[55] M. Raynaud, H. Zhang, K. Louis, V. Goutaudier, J. Wang, Q. Dubourg, Y. Wei, Z. Demir, C. Debiais, O. Aubert, Y. Bouatou, C. Lefaucheur, P. Jabre, L. Liu, C. Wang, X. Jouven, P. Reese, J.-P. Empana, and A. Loupy. COVID-19-Related

Medical Research: a Meta-research & Critical Appraisal. *BMC Med. Res. Methodology*, 21(1), Jan. 2021. doi: 10.1186/s12874-020-01190-w.

[56] J. T. Reese, D. Unni, T. J. Callahan, L. Cappelletti, V. Ravanmehr, S. Carbon, K. A. Shefchek, B. M. Good, J. P. Balhoff, T. Fontana, H. Blau, N. Matentzoglu, N. L. Harris, M. C. Munoz-Torres, M. A. Haendel, P. N. Robinson, M. P. Joachimiak, and C. J. Mungall. KG-COVID-19: A Framework to Produce Customized Knowledge Graphs for COVID-19 Response. *Patterns*, 2(1):100155, Jan. 2021. doi: 10.1016/j.patter.2020.100155.

[57] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing*. Ass'n for Computational Linguistics, Nov. 2019. URL `https://arxiv.org/abs/1908.10084`.

[58] A. Rovetta. Reliability of Google Trends: Analysis of the Limits & Potential of Web Infoveillance During COVID-19 Pandemic & for Future Research. *medRxiv*, 2021. doi: 10.1101/2020.12.29.20248969.

[59] A. Sancheti, K. Chawla, and G. Verma. LynyrdSkynyrd at WNUT-2020 Task 2: Semi-Supervised Learning for Identification of Informative COVID-19 English Tweets. pages 444–449, Nov. 2020. doi: 10.18653/v1/2020.wnut-1.65.

[60] A. C. Sanders, R. C. White, L. S. Severson, R. Ma, R. McQueen, P. H. C. Alcânatara, Y. Zhang, J. S. Erickson, and K. P. Bennett. Unmasking the Conversation on Masks: Natural Language Processing for Topical Sentiment Analysis of COVIDs-19 Twitter Discourse. *medRxiv*, 2020. doi: 10.1101/2020.08.28.20183863.

[61] C. Savvides and R. Siegel. Asymptomatic & Presymptomatic Transmission of SARS-CoV-2: Systematic Review. Jun. 2020. doi: 10.1101/2020.06.11.20129072.

[62] H. J. Schünemann, N. Santesso, G. E. Vist, C. Cuello, T. Lotfi, S. Flottorp, M. Davoli, R. Mustafa, J. J. Meerpohl, P. Alonso-Coello, and E. A. Aklh. Using GRADE in Situations of Emergencies & Urgencies: Certainty in Evidence & Recommendations Matters During the COVID-19 Pandemic, Now More Than Ever & No Matter What. *J. Clin. Epidemiol.*, Nov. 2020. doi: doi:10.1016/j.jclinepi.2020.05.030.

[63] S. SeyedAlinaghi, S. Oliaei, S. Kianzad, A. M. Afsahi, M. MohsseniPour, A. Barzegary, P. Mirzapour, F. Behnezhad, T. Noori, E. Mehraeen, O. Dadras, F. Voltarelli, and J.-M. Sabatier. Reinfection Risk of Novel Coronavirus (COVID-19): a Systematic Review of Current Evidence. *World J. Virology*, 9(5):79–90, Dec. 2020. doi: 10.5501/wjv.v9.i5.79.

[64] J. Shuja, E. Alanazi, W. Alasmary, and A. Alashaikh. COVID-19 Open Source Data Sets: a Comprehensive Survey. *Applied Intel.*, 51(3):1296–1325, Sept. 2020. doi: 10.1007/s10489-020-01862-6.

[65] K. Soares-Weiser, T. Lasserson, K. J. Jorgensen, S. Woloshin, L. Bero, M. D. Brown, and B. Fischhoff. Policy Makers Must Act on Incomplete Evidence in Responding to COVID-19. *Cochrane Database of Systematic Revs.*, Nov. 2020. doi: 10.1002/14651858.ed000149.

[66] A. Spangher, N. Peng, J. May, and E. Ferrara. Enabling Low-Resource Transfer Learning across COVID-19 Corpora by Combining Event-Extraction & Co-Training. In *Proc. of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, Jul. 2020. Ass'n for Computational Ling. URL `https://www.aclweb.org/anthology/2020.nlpCOVID19-acl.4`.

[67] M. Udow-Phillips and P. M. Lantz. Trust in Public Health Is Essential Amid the COVID-19 Pandemic, Jun. 2020.

[68] K. Verspoor, S. Šuster, Y. Otmakhova, S. Mendis, Z. Zhai, B. Fang, J. H. Lau, T. Baldwin, A. J. Yepes, and D. Martinez. COVID-SEE: Scientific Evidence Explorer for COVID-19 Related Research. 2020. URL `https://arxiv.org/abs/2008.07880`.

[69] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi. Fact or Fiction: Verifying Scientific Claims. 2020. URL `https://arxiv.org/abs/2004.14974`.

[70] L. L. Wang and K. Lo. Text Mining Approaches for Dealing With the Rapidly Expanding Literature on COVID-19. *Briefings in Bioinformatics*, 22(2):781–799, Dec. 2020. doi: 10.1093/bib/bbaa296.

[71] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier. CORD-19: The COVID-19 Open Research Dataset. In *Proc. of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, Jul. 2020. Ass'n for Computational Linguistics.

[72] W. Wang, Y. Wang, X. Zhang, X. Jia, Y. Li, and S. Dang. Using WeChat, a Chinese Social Media App, for Early Detection of the COVID-19 Outbreak in December 2019: Retrospective Study. *JMIR mHealth and uHealth*, 8(10):e19589, Oct. 2020. doi: 10.2196/19589.

[73] T. Weissgerber, N. Riedel, H. Kilicoglu, C. Labbé, P. Eckmann, G. ter Riet, J. Byrne, G. Cabanac, A. Capes-Davis, B. Favier, S. Saladi, P. Grabitz, A. Bannach-Brown, R. Schulz, S. McCann, R. Bernard, and A. Bandrowski. Automated Screening of COVID-19 Preprints: Can We Help Authors to Improve Transparency & Reproducibility? *Nature Med.*, 27(1):6–7, Jan. 2021. doi: 10.1038/s41591-020-01203-7.

[74] K. A. Whitmore, K. B. Laupland, C. M. Vincent, F. A. Edwards, and M. C. Reade. Changes in Medical Scientific Publication Associated With the COVID-19 Pandemic, Nov. 2020.

[75] H. Xia, W. An, J. Li, and Z. J. Zhang. Outlier Knowledge Management for Extreme Public Health Events: Understanding Public Opinions About COVID-19 Based on Microblog Data. *Socio-Economic Planning Sciences*, page 100941, Sept. 2020. doi: 10.1016/j.seps.2020.100941.