Disaster Medicine and Public Health Preparedness

www.cambridge.org/dmp

Concepts in Disaster Medicine

Cite this article: Dobolyi K, Sieniawski GP, Dobolyi D, Goldfrank J, Hampel-Arias Z. Hindsight2020: Characterizing uncertainty in the COVID-19 scientific literature. *Disaster Med Public Health Prep*. **17**(e437), 1–8. doi: https://doi.org/10.1017/dmp.2023.82.

Keywords:
SARS-CoV-2; uncertainty; natural language processing; public health; pandemics

Corresponding author:
Kinga Dobolyi;
Email: kinga@gwu.edu

Hindsight2020: Characterizing Uncertainty in the COVID-19 Scientific Literature

Kinga Dobolyi PhD[1], George P. Sieniawski MS[2], David Dobolyi PhD[3], Joseph Goldfrank PhD[1] and Zigfried Hampel-Arias PhD[4]

[1]George Washington University, Department of Computer Science, Washington, DC, USA; [2]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; [3]University of Notre Dame, Indiana, USA and [4]Los Alamos National Laboratory, Los Alamos, New Mexico, USA

Abstract

Following emerging, re-emerging, and endemic pathogen outbreaks, the rush to publish and the risk of data misrepresentation, misinterpretation, and even misinformation puts an even greater onus on methodological rigor, which includes revisiting initial assumptions as new evidence becomes available. This study sought to understand how and when early evidence emerges and evolves when addressing different types of recurring pathogen-related questions. By applying claim-matching by means of deep learning Natural Language Processing (NLP) of coronavirus disease 2019 (COVID-19) scientific literature against a set of expert-curated evidence, patterns in timing across different COVID-19 questions-and-answers were identified, to build a framework for characterizing uncertainty in emerging infectious disease (EID) research over time. COVID-19 was chosen as a use case for this framework given the large and accessible datasets curated for scientists during the beginning of the pandemic. Timing patterns in reliably answering broad COVID-19 questions often do not align with general publication patterns, but early expert-curated evidence was generally stable. Because instability in answers often occurred within the first 2 to 6 mo for specific COVID-19 topics, public health officials could apply more conservative policies at the start of future pandemics, to be revised as evidence stabilizes.

Introduction

Although coronavirus disease2019 (COVID-19) prompted a rapid surge in scientific research activity, several questions remained unsettled even a year and a half after the World Health Organization (WHO)'s pandemic declaration in March 2020. For instance, it was too soon to characterize long-term disease sequelae even in early 2021 (a year later), and immunity duration and the risks of breakthrough infections[1] were not immediately obvious. Various transmissibility-related questions divided researchers in early 2020,[2] resulting in diminished trust in mask guidance.[3] Because communicating uncertainty about emerging infectious disease outbreaks is inherently difficult, scientists and policy-makers typically use a diverse set of approaches for distilling insights, acknowledging evidence gaps, updating public health guidance, and adjusting mitigation measures.[4,5]

Among these approaches is the Department of Homeland Security (DHS) Master Question List (MQL), discussed below, which outlines known unknowns about novel pathogens.[6] In addition to the MQL, related approaches like Grading of Recommendations Assessment, Development and Evaluation (GRADE)[7] can help global health organizations such as the WHO formulate outbreak response strategies over a realistic range of time frames, while considering varying levels of evidence quality. Some data, like those involving randomized controlled trials of vaccines, take a while to gather sufficient data to resolve. Others, like those about decontamination, require relatively modest time investments to answer, both for initial guidance purposes and over longer timeframes. To compound this challenge, vanity articles[8] and opinion pieces lacking novel results can also overshadow bona fide hypothesis development within the onset-stage pandemic literature.[9]

Motivation

This study examines the evolution of useful information on novel pathogens within scientific literature. Its goal was 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 collection practice, and interdisciplinary citation networks, consistent with GRADE.[7] To do so, human-curated evidence was traced through scientific publications on SARS-CoV-2 over time to generate timelines of when questions are addressed, and how answers evolve.

By applying claim-matching, by means of Natural Language Processing (NLP), to the roughly 600 sentences of evidence DHS cited in January 2021 in reply to their 16 MQL questions,[6] this study sought to match each of these hundreds of sentences to similar and related claims in a snapshot of the COVID-19 Open Research Dataset (CORD-19) corpus[10] of 13 million sentences mined from the scientific literature on severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) during the same time-frame (March 2020 through January 2021). By then analyzing the timing and uncertainty of new evidence over time, this preliminary framework provides a foundation for global health experts and policy-makers during the onset phase of emerging pathogen outbreaks. The aim is to characterize when recurring questions about different types of diseases might begin to yield reliable answers within the first year of a new outbreak.

Specifically, this study ranked sentence similarity in terms of how closely aligned 2 sentences are from the standpoint of research claim-matching, that is: to find which academic papers contained sentences similar in spirit to the original DHS evidence. Statements describing the same phenomenon, even if they arrived at opposing research conclusions, are included. With the goal of characterizing when different questions can be reliably answered by means of scientific publications, this work, therefore, investigated the following core hypotheses with respect to the DHS Master Question List:

- Hypothesis 1: Different COVID-19 questions are answered at different times
- Hypothesis 2: Contradictory evidence appears after initial claims
- Hypothesis 3: Evolution in (un)certainty of claims varies across questions

Related Work: Pandemic Uncertainty and Risk Communications

Although the COVID-19 literature has grown exponentially since January 2020, only 20% of preprints on COVID-19 later appeared in peer-reviewed journals.[11] 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.[12–15] In addition, between the start of the pandemic and May 2020, the majority of published material did not contain original data (eg, opinion pieces).[13] Of interest, COVID-19 publication growth in the first year appears to have reached its apogee in May 2020 and subsequently trended downward to November 2020,[16] potentially indicating less conjecture and diminished levels of vanity publishing. Of note, we chose to study COVID-19 because of the timely, public availability of such large, curated datasets like CORD-19; we are not aware of similar resources for other infectious diseases.

However uncomfortable for policy-makers, acknowledging the uncertainty associated with scientific evidence is a source of credibility and a means of retaining public trust.[17] The converse also appears to be true: overconfidence can backfire.[18] Additionally, information overload about pandemics like COVID-19 (which often involves conflicting evidence)[19] is a significant risk. 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[5] when policy experts had to extrapolate from prior experiences with other pathogens.

Methods

The Department of Homeland Security (DHS) updates its Master Question List (MQL)[6] citations on an ongoing basis, and this study obtained a publicly-available update from December 21, 2020, which provided expert-curated evidence to answer 1 of 16 questions. Almost 600 sentences of evidence were provided to answer these 16 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]* are listed as 1 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 the ground truth claims for this work, and each may be associated with 1 or more cited academic articles, trusted publications, news sources, and other materials in the COVID-19 MQL.

Construction of HindSight2020 Dataset

Matching MQL ground-truth claims against evidence in the CORD-19 dataset of academic articles

This research constructed its claim-matching dataset for SARS-CoV-2 academic sentence pairs by using a snapshot of the large CORD-19 corpora[10] obtained on January 4, 2021. In its construction, CORD-19 papers were sourced from PubMedCentral, PubMed, the WHO's COVID19 database, and preprint servers bioRxiv, medRxiv, and arXiv, collecting papers that contained specific SARS and/or MERS keywords. In April 2020, roughly half of these papers were from the field of Medicine, a third from Biology, and the most common subfields were Virology (26%), Immunology (14%), and Surgery (14%); however, these ratios may have evolved as time progressed. Sentences from both article abstracts and bodies were included for all articles and/or preprints that appeared in the ≈144,000 article parses available. After filtering out articles predating 2020 (as the CORD-19 dataset includes articles on diseases potentially related to COVID-19, such as Ebolavirus and influenza), over 13 million sentences were obtained against which to compare the ≈600 DHS claims.

Filtering the CORD-19 dataset of academic articles

Next, these ≈13 million sentences from CORD-19 were filtered into a much smaller subset of potentially matching claims for each of the ≈600 ground truth sentences from the DHS MQL , as shown in Figure 1. To do so, Named Entity Recognition (NER) using spaCy's[20] pretrained en core sci sm model was applied to identify relevant keywords in each of the DHS sentences. For almost all DHS sentences, the CORD-19 sentences were filtered to those with at least 3 uncased keyword matches between the 2 sentences, which reduced computation time to approximately a day.

Mining matched claims using SBERT

Once each of the ≈600 MQL claims had its own subset of keyword-matched sentences to match against from the CORD-19 dataset, SBERT,[21] a deep learning NLP model often used to detect Semantic Textual Similarity between sentence pairs, was used to perform claim-matching. SBERT was configured to provide up to the top 10

https://doi.org/10.1017/dmp.2023.82 Published online by Cambridge University Press

matching sentenced from CORD-19 for each of the ≈600 ground truth claims. In the results, this study was often able to trace back nonacademic DHS citations to older academic sources. In cases where DHS citations existed in the CORD-19 database, the claim-matching approach (described in this section) was able to directly match the cited article approximately 20% of the time; for the remaining 80%, expert annotators were able to evaluate similarity manually.

While such a low initial match might seem like a disappointment, in this case it was desirable, as the goal was not to prove this study could replicate the original paper citations, but rather to trace the evolution of those claims through subsequent research, either with similar research results in another study, or a mention in a related work section. Often, DHS-cited claims were paraphrased or summarized to the point of an inability to detect any meaningfully related sentences (such as *Reinfection is possible.*); this occurred in 4% of the DHS sentences. A total of 346 of the ≈600 ground truth claims had at least 1 study cited directly by DHS found in the CORD-19 dataset (although not necessarily matched by the SBERT algorithm this study used). Remaining DHS citations were frequently non-academic articles (eg, news sources like *Reuters*, government surveillance reports, or press releases).

Expert annotation of HindSight2020 matched claim pairs

Next, the quality of the SBERT-matched claim-sentence pairs were evaluated using human experts, to ensure that further analysis was meaningful, using the scoring in Table 1. Sentence-scoring is a resource-intensive, expensive, and arguably subjective process, even for experts. This study relied on a single expert annotator to rate all 5814 sentence pairs initially; this bioinformatician had extensive background in the COVID-19 literature, following the academic articles closely through this work's timeframe. Because the goal was not to evaluate the quality of that research literature, but rather, to determine whether the 2 sentences described the same phenomenon, the task did not require deep biological expertise. Table 2 is an example of 3 potential matches for the DHS claim *This could mean that 5–10% of individuals undergoing a 14-d quarantine are still infectious at the end*, along with their expert-judged research-similarity rating. This study suggests researchers consider these annotations (available at https://github.com/IQTLabs/hindsight2020) as more of a continuous range, rather than a strict binning. Finally, the DHS evidence for forecasting models (DHS Question 16), yielded low-quality matches by means of SBERT, possibly because the DHS language here was a bullet list of incomplete sentences, which could be more difficult to match for similarity using the algorithm.

Results

To analyze the timing patterns of evidence presentation and collection for the DHS MQL, the annotated sentence pairs above were filtered to include only the set of sentences the annotator labeled as *yes, definitely*, or *perhaps* (sentences describing the same phenomenon or conceptually related work). Sentences that appeared to be duplicative citations or references to earlier studies were also manually filtered out to minimize the impact of these types of sentences. The analysis below presents findings on this subset of evidence (ie, close matches of original research to the DHS claims).

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 2 sentences are not describing the same things

Table 2. Annotator rating scale examples

DHS sentence	This could mean that 5-10% of individuals undergoing a 14-d quarantine are still infectious at the end.
Rating	CORD-19 sentence match to DHS sentence above
Yes, definitely	Notably, even a 14-d quarantine period does not eliminate the risk of individuals spending time infectious after release.
Perhaps	Under a 5-d quarantine period, around 6.8% of infected arrivals are released while highly infectious.
No	However, if this study assumes anyone who is infected by the virus will be under quarantine 7 d 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.

Hypothesis 1: Different Questions Are Answered at Different Times

The first hypothesis was that high-quality evidence in CORD-19 for the 16 DHS MQL questions would emerge at different times, and not match the pattern of exponential publication growth from January through May 2020, followed by a slow but steady decline through the rest of that calendar year. All the statements obtained from the filtering process described above are plotted in Figure 2. While one would expect evidence involving vaccines and protective immunity to accumulate later, this study revealed that even questions around personal protective equipment (PPE), transmissibility, clinical diagnosis, and environmental stability continued to occupy researchers for months, through the second half of 2020. While it is possible statements that referred to earlier research were not filtered out (as the natural language in academic articles may refer to another study obliquely or omit a citation entirely), this study also showed in hindsight that researchers revised many answers to the MQL questions in light of subsequent findings, as discussed in section 4.4.

The timing of DHS claims against the timing of close matching evidence (as above, but with the added restriction of including only *yes, definitely*, labels) was compared next in the CORD-19 dataset. This study hypothesized that 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, except for *Decontamination*, as shown in Figure 3. Obtaining answers to the MQLs as soon as possible is the goal of effective public health policy-making, as it enables timelier crisis response and resource allocation, ultimately saving lives and minimizing the impact of emerging disease outbreaks.[4] For SARS-CoV-2, it appears the DHS MQL compilers were able to identify meaningful answers to these questions early and often across a range of pandemic-related issues.

https://doi.org/10.1017/dmp.2023.82 Published online by Cambridge University Press

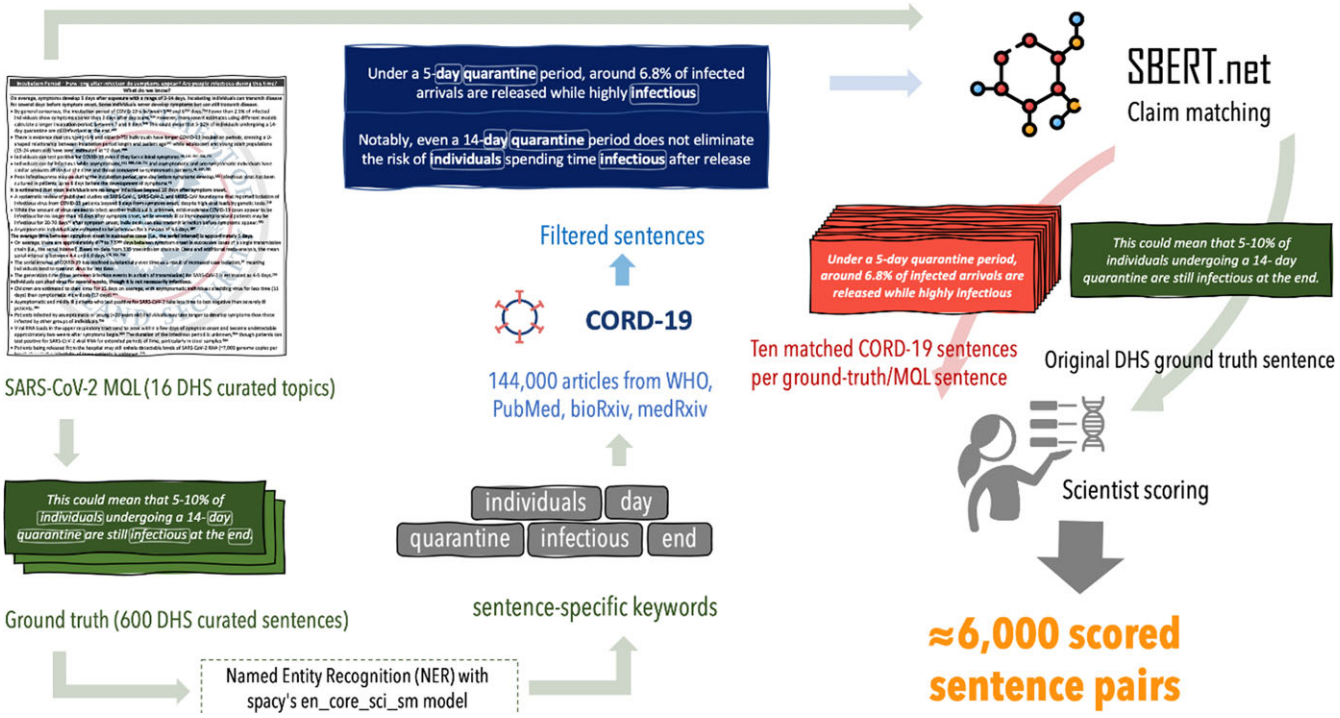

Figure 1. Flowchart of construction of our dataset.

Hypothesis 2: Contradictory Evidence Appears After Initial Claims

Early MQL answers that do not change later are especially valuable. To measure how often this occurred, entailment vs. contradiction was analyzed within DHS-CORD-19 sentence pairs labeled as *yes, definitely,* or *perhaps* by the expert annotator. Using the MedNLIs glove-bio-asq-mimic model from BioASQ,[22] this study predicted whether the matched claim was an entailment or contradiction of the DHS claim. This automated approach generated many false positives, so manual review of each pair labeled as an apparent contradiction was required, arriving at ≈40 actual contradictions in the dataset of 5814 sentence pairs, plotted in Figure 4. Of the ≈40 contradictions that had the study cited by DHS in the CORD-19 dataset (and, therefore, contained publication date), 8 were cases where the DHS evidence cited the new research; very rarely did the original DHS conclusions later change.

Most contradictions could be found within ≈10 to ≈30 wk from the 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 anakinra, 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 around aerosol transmission and presymptomatic 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.

Hypothesis 3: Evolution in (Un)Certainty of Claims Varies Across Questions

These results show the research published early during a pandemic, despite all of its limitations, can be successfully curated by humans into a set of early, actionable evidence, with the caveat that some MQL answers are more prone to revision than others. Next, this study wanted to investigate the language scientists use within their publications to communicate this uncertainty: does it change over time? The *Linguistic Uncertainty Classifier Interface* by Vincze et al.[23,24] was used to label the language of every DHS and CORD-19 sentence of evidence as either certain or uncertain. Each DHS-CORD-19 sentence pair over time was then plotted, using the CORD-19 paper date, to show how the MQL evidence to CORD-19 evidence transitioned over time: either *certain to certain (C-C)*, *certain to uncertain (C-U)*, *uncertain to certain (U-C)*, or *uncertain to uncertain (U-U)* over time. Only sentence pairs below where the CORD-19 evidence came later than the MQL evidence (when this study had a date available for the latter) were graphed. Each tile in Figure 5 represents the evolution of uncertainty over time, per question, between the earlier MQL claim and the matching CORD-19 evidence. No points could be plotted for Question 9 (Vaccines) and Question 16 (Forecasting) due to a lack of matching CORD-19 papers or timestamped citations from DHS.

Overall, many answers seemed to express certainty throughout the pandemic more often than not, including incubation period, clinical presentation, environmental stability, and decontamination. Meanwhile, transmissibility, medical treatments, non-pharmaceutical interventions (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,

https://doi.org/10.1017/dmp.2023.82 Published online by Cambridge University Press

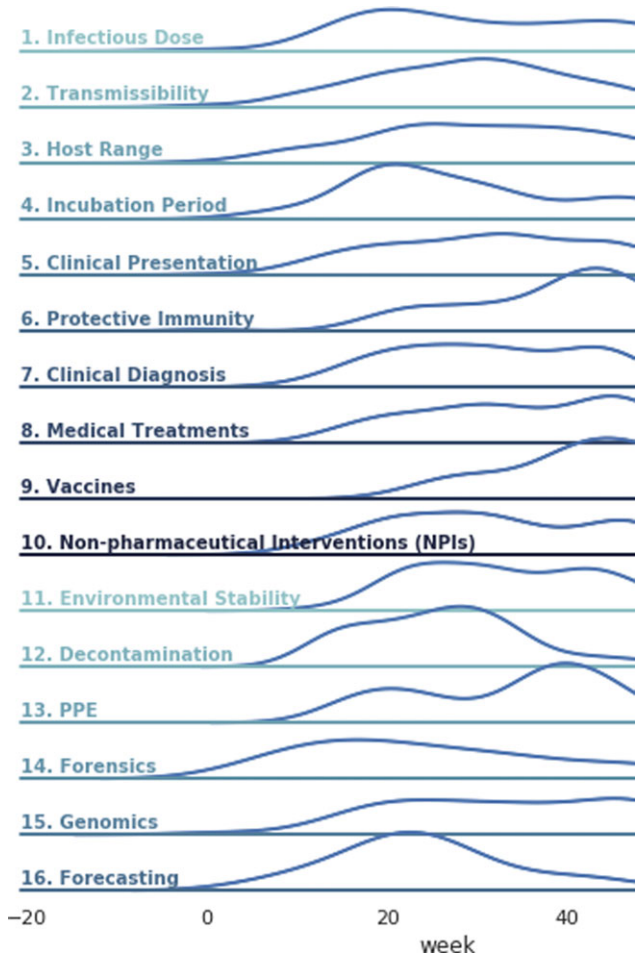

Figure 2. Timing patterns of CORD-19 matches of original research from start week of pandemic (week 0, March 2020) through January 2021. The y-axis lists the density, over time, of CORD-19 evidence sentences that matched the DHS-cited evidence sentences for each of the 16 questions.

the evolution this investigation was 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 3 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.[25]

Limitations

There are several limitations to the work presented here, including limitations of claim-matching.

For example, this study may have missed matching claims in CORD-19 due to the MQL ground truth sentence being a paraphrase, or SBERT, and/or NER tools potentially not recognizing synonyms – eg, a *migraine* might be equivalent for our purposes to a *headache*. These and similar limitations open the possibility of false negatives in the automated claim-matching approaches presented. Given the subjectivity of our expert annotations, it is also possible that there are additional false positive and false negative matches. Another limitation was the reliance on a single expert annotator to decide the quality of SBERT matches, due to the expense of this task. Although the preliminary results indicated that when the same expert re-rated samples of the 10 matches for 10 arbitrary DHS claims (100 sentence-pairs total), they arrived at the same rating 95% of the time, a separate annotator agreed with the expert 85% of the time. A study to formally measure intra- and inter-annotator agreement will be conducted in an ongoing follow-up to this work.

In addition, the DHS MQL is a living document, and this study only analyzed ground truths for a single snapshot in time; the reason its evidence was found to be so stable may be that it had already undergone revision. However, during the (un)certainty and contradiction analyses, this study 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; future work could explore such evolution of weekly updates to the MQL. Finally, this study only examined the timelines for evidence collection across DHS questions for SARS-CoV-2 in this work. The timelines for other pathogens, such as monkeypox, and outbreak scenarios may differ substantially.

Discussion

Given the deluge of academic preprints and peer-reviewed papers on SARS-CoV-2 in 2020,[11] this study sought to determine if it were 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, the authors were curious what happened to early evidence mined to answer other types of questions (e.g., about clinical presentation, transmission, and decontamination).

Overall, this study found that most early human-curated evidence DHS compiled into the MQL was highly reliable and stable over time. When newer evidence contradicted original conclusions, this generally happened within 2 to 6 mo of the start of the pandemic. Therefore, it seems the highest risk of evidence changing occurs in the first 6 mo 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 6-mo 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 the framework of uncertainty, this study found this shift correlated with contradictions in the literature.

To summarize, the core pillars of the framework this study thus proposes are:

1. There exist important, unanswered questions at the start of novel infectious disease outbreaks (such as the 16 questions defined by the DHS MQL);
2. The timing of reliable answers to each of these questions varies based on the nature of the question;
3. Answers to some questions, more than others, may be subject to revision as time goes on;
4. Therefore, this study recommends building public health policy under the assumption that recommendations need to be revisited within the first 6 mo of a new outbreak, and that more conservative initial measures may be appropriate as these can be relaxed once more information is expected to emerge in such a 6-mo timeframe.

https://doi.org/10.1017/dmp.2023.82 Published online by Cambridge University Press

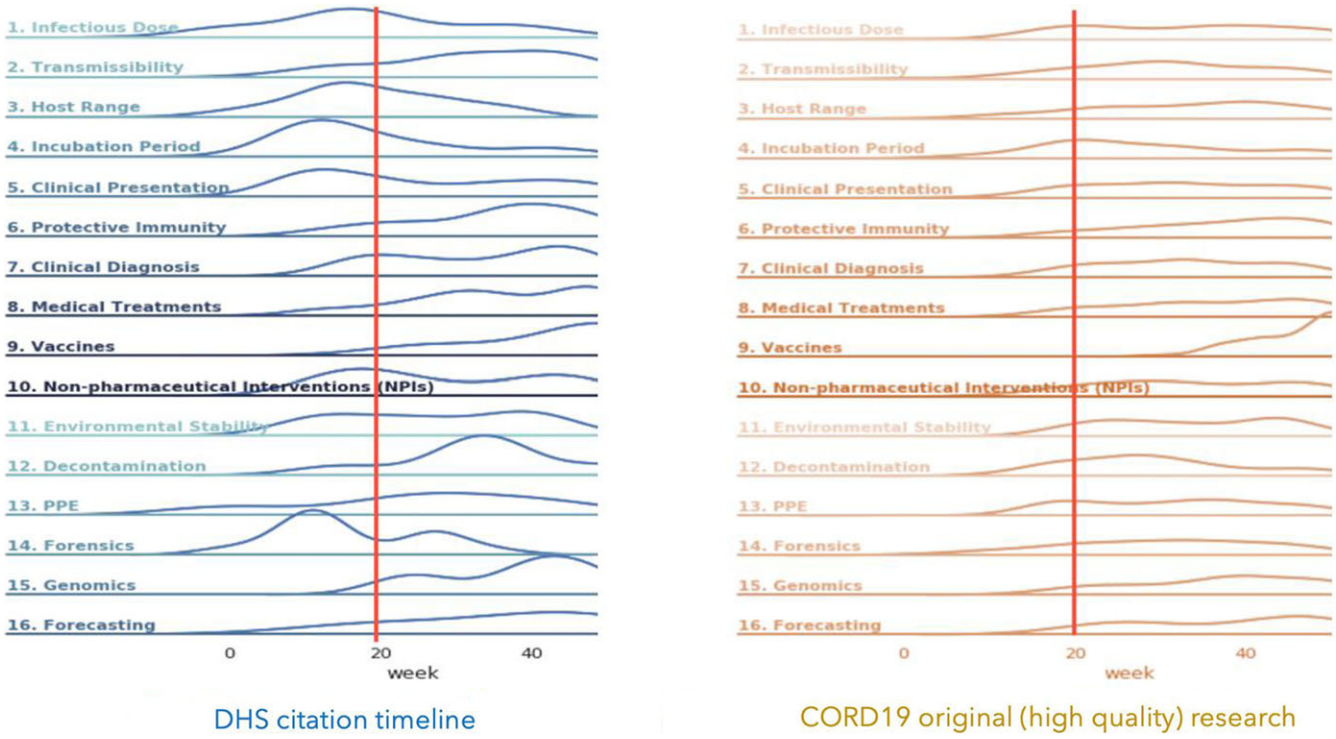

Figure 3. Timing patterns of DHS evidence vs. CORD-19 close matches of original research from start week of pandemic (week 0, March 2020) through January 2021. The y-axis lists the density, over time, of evidence sentences (either DHS or CORD-19 matched) for each of the 16 questions.

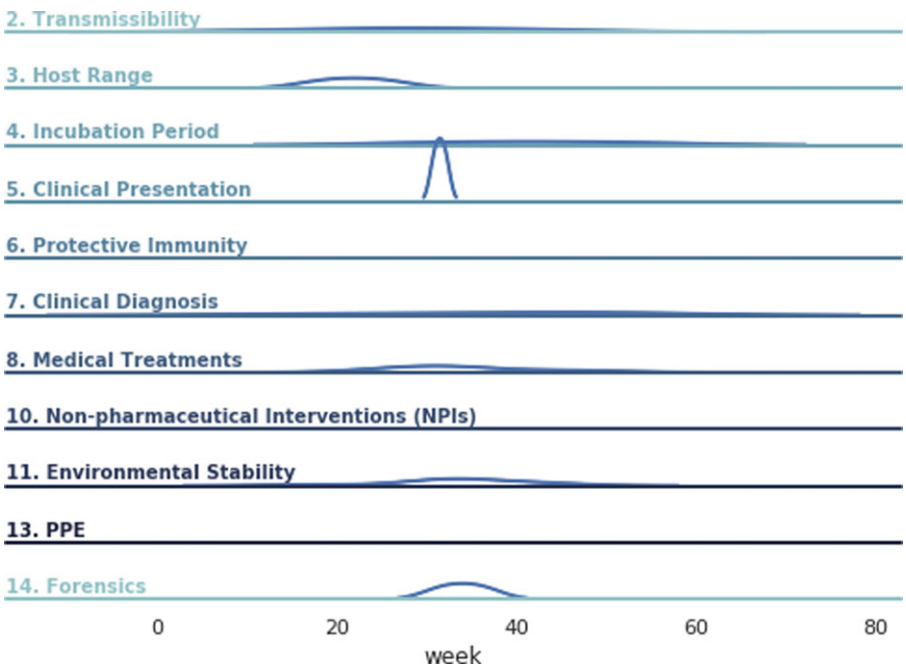

Figure 4. Timing patterns of results contradicting DHS claims from start week of pandemic (week 0, March 2020) through January 2021. The y-axis lists the density, over time, of DHS-cited evidence sentences, that were contradictions with earlier results, for each of the 16 questions.

This 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.

https://doi.org/10.1017/dmp.2023.82 Published online by Cambridge University Press

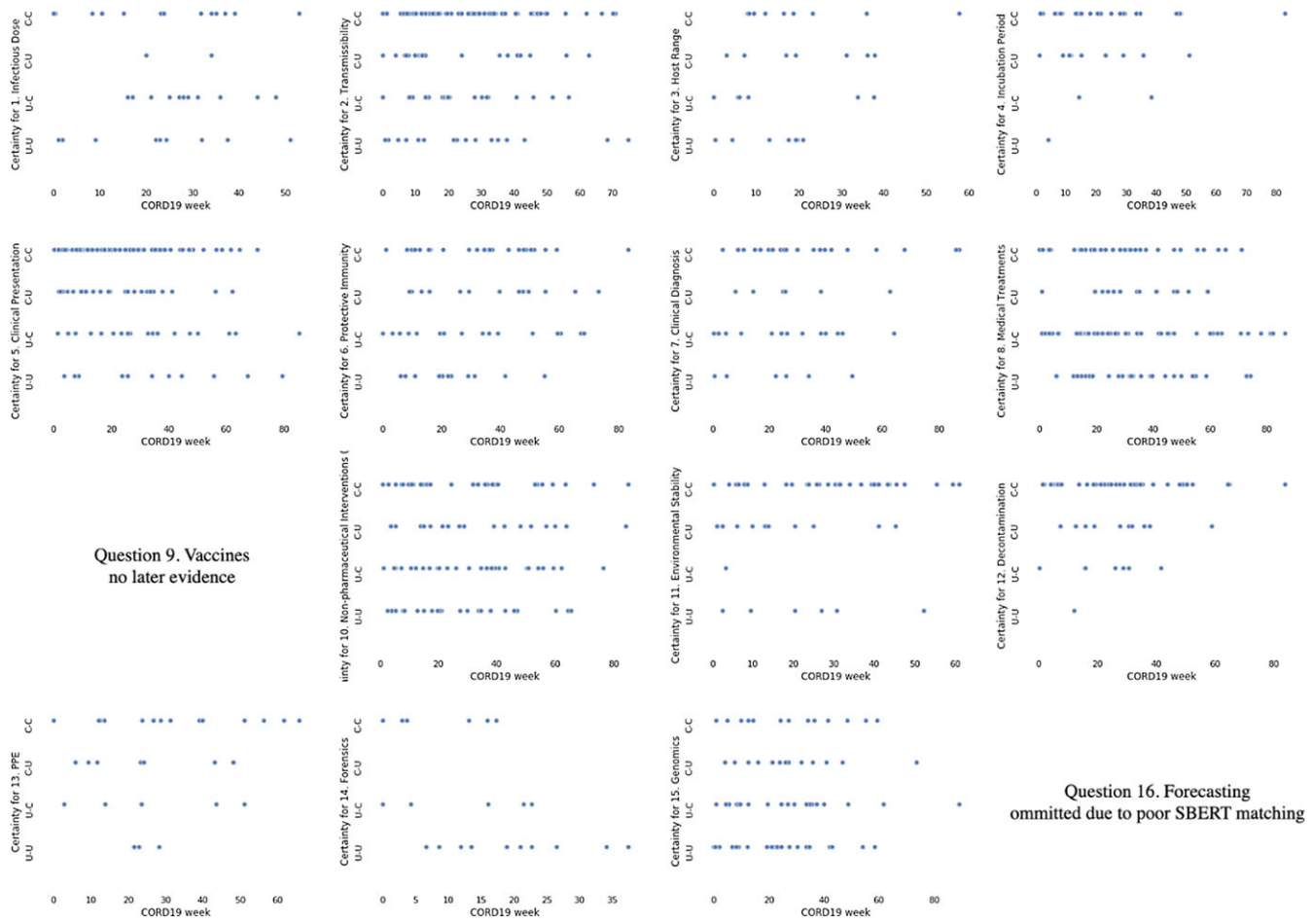

Figure 5. Uncertainty evolution over time between MQL claims and matching CORD-19 evidence from a later date, by week from pandemic start. The y-axis shows the potential evolution (or lack thereof) between C-C, C-U, U-C, or U-U as defined above.

While the GRADE[7] rating system can assign a quality rating to evidence, such that we would expect higher-quality evidence to remain more stable, such high-quality studies with proper sample sizes and experimental conditions are often difficult, if not impossible, to obtain early on in pandemics. Therefore, we offer a complimentary approach that seeks to identify what evidence is likely to change during the first 6 mo of a pandemic, assuming that much of the evidence will not be high-quality during an emerging disease.

Conclusions

This study created the foundations for a framework to characterize uncertainty around evidence in the context of academic articles on emerging infectious diseases. The goal was to understand how the evidence used to answer different pandemic-related questions 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 postpeak 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 risk communications may be most critical.

Acknowledgments. We thank Dylan George, Dan Hanfling MD, Kevin O'Connell MD, 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, where Nina Lopatina was also a contributor to that claim-matching repository as the project lead.

Author contribution. G.S. and K.D. conceived of the presented idea. K.D. designed, coded (with JG), and performed the experiments. K.D. processed the experimental data, performed the analysis, drafted the manuscript, and designed the figures and analyzed the data. K.D. and J.G. were the primary and secondary annotators of the dataset, respectively. D.D. and J.G. helped design the experimental annotation. K.D. and G.S. wrote the study with input from all the authors. Z.H.A. suggested that citations be filtered out from matched sentences in our experimental design.[1]Research conducted while at IQT Labs: Kinga Dobolyi, George P. Sieniawski, Zigfried Hampel-Arias.

Competing interests. We are reporting that IQT Labs (who sponsored this research, in part) by means of their parent organization, In-Q-Tel, maintains a professional relationship (including funding) with government entities that may be affected by these findings.

Statement of IRB approval or exemption from full review. As this study only analyzed existing natural language datasets from DHS and CORD-19, and did not involve human nor biological experimentation of any sort, this work does not fall under a category that requires IRB approval. The 2 annotators mentioned were authors of this study (Kinga Dobolyi and Joseph Goldfrank).

https://doi.org/10.1017/dmp.2023.82 Published online by Cambridge University Press

References

1. **SeyedAlinaghi S, Oliaei S, Kianzad S, *et al.*** Reinfection risk of novel coronavirus (COVID-19): a systematic review of current evidence. World J Virol. 2020;9(5):79-90. doi: 10.5501/wjv.v9.i5.79.
2. **Savvides C, Siegel R.** Asymptomatic and presymptomatic transmission of SARS-CoV-2: a systematic review. 2020. doi: 10.1101/2020.06.11.20129072
3. **Udow-Phillips M, Lantz PM.** Trust in public health is essential amid the COVID-19 pandemic. *J Hosp Med.* 2020;15(7):431-433. doi: 10.12788/jhm.3474
4. **Berger L, Berger N, Bosetti V, *et al.*** Rational policymaking during a pandemic. *Proc Natl Acad Sci USA*. 2021;118(4):e2012704118. doi: 10.1073/pnas.2012704118.
5. **Soares-Weiser K, Lasserson T, Juhl Jorgensen K, *et al.*** Policy makers must act on incomplete evidence in responding to COVID-19. Cochrane Database Syst Rev. 2020;11: ED000149. doi: 10.1002/14651858.ED000149
6. **US Department of Homeland Security**. Master question list for COVID-19 (caused by SARS-CoV-2). Accessed December 21, 2020. https://www.dhs.gov/publication/st-master-question-list-COVID-19, 2022
7. **Schünemann HJ, Santesso N, Vist GE, *et al.*** Using GRADE in situations of emergencies and urgencies: certainty in evidence and recommendations matters during the COVID-19 pandemic, now more than ever and no matter what. *J Clin Epidemiol.* 2020;127:202-207. doi: 10.1016/j.jclinepi.2020.05.030
8. **Jalali R, Hosseinian-Far A, Mohammadi M.** Contradictions in the promotion of publishing academic and scientific journal articles, and the inability to cope with the new coronavirus (COVID-19). *Antimicrob Resist Infect Control.* 2021;10(1):10. doi: 10.1186/s13756-021-00884-0
9. **Odone A, Galea S, Stuckler D, *et al.*** The first 10 000 COVID-19 papers in perspective: are we publishing what we should be publishing? *Eur J Public Health.* 2020;30(5):849-850. doi: 10.1093/eurpub/ckaa170
10. **Wang LL, Lo K, Chandrasekhar Y, *et al.*** CORD-19: The COVID-19 open research dataset. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. 2020; arXiv:2004.10706v4. Association for Computational Linguistics.
11. **Älgå A, Eriksson O, Nordberg M.** The development of preprints during the COVID-19 pandemic. *J Intern Med.* 2021;290(2):480-483. doi: 10.1111/joim.13240
12. **Elgendy IY, Nimri N, Barakat AF, *et al.*** A systematic bias assessment of top-cited full-length original clinical investigations related to COVID-19. *Eur J Intern Med.* 2021;86:104-106. doi: 10.1016/j.ejim.2021.01.018
13. **Raynaud M, Zhang H, Louis K, *et al.*** COVID-19-related medical research: a meta-research and critical appraisal. *BMC Med Res Methodol.* 2021;21(1):1. doi: 10.1186/s12874-020-01190-w
14. **Whitmore KA, Laupland KB, Vincent CM, *et al.*** Changes in medical scientific publication associated with the COVID-19 pandemic. *Med J Australia.* 2020;213(11):496-499. doi: 10.5694/mja2.50855
15. **Palayew A, Norgaard O, Safreed-Harmon K, *et al.*** Pandemic publishing poses a new COVID-19 challenge. *Nat Hum Behav.* 2020;4(7):666-669. doi: 10.1038/s41562-020-0911-0
16. **Kang M, Gurbani SS, Kempker JA.** The published scientific literature on COVID-19: an analysis of Pubmed abstracts. *J Med Sys.* 2020;45(1):3. doi: 10.1007/s10916-020-01678-4
17. **Fiske ST, Dupree C.** Gaining trust as well as respect in communicating to motivated audiences about science topics. *Proc Natl Acad Sci USA.* 2014;111(Suppl 4):13593-13597. doi: 10.1073/pnas.1317505111
18. **Pearce W.** Trouble in the trough: how uncertainties were downplayed in the UK's science advice on COVID-19. *Humanit Soc Sci Commun.* 2020. doi: 10.1057/s41599-020-00612-w
19. **Mohammed M, Sha'aban A, Jatau AI, *et al.*** Assessment of COVID-19 information overload among the general public. *J Racial Ethn Health Disparities.* 2021;9(1):184-192. doi: 10.1007/s40615-020-00942-0
20. **Montani I, Honnibal M, Van Landeghem S, *et al.*** spaCy: industrial-strength natural language processing in Python. 2020. Accessed May 29, 2023. https://zenodo.org/record/4021943
21. **Reimers N, Gurevych I.** Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2019. doi: 10.48550/arXiv.1908.10084
22. **BioASQ.org**. BioASQ releases continuous space word vectors obtained by Applying Word2Vec to PubMed Abstracts. Accessed December 3, 2021. http://bioasq.org/news/bioasq-releases-continuous-space-word-vectors-obtained-applying-word2vec-pubmed-abstracts
23. **Meyers B.** meyersbs/uncertainty. Installation & usage. Accessed May 29, 2023. https://github.com/meyersbs/uncertainty/wiki/installation-&-usage.
24. **Vincze, V.** Uncertainty Detection in Natural Language Texts. *University of Szeged.* 2014. doi: 10.14232/phd.2291
25. **Bero L, Lawrence R, Leslie L, *et al.*** Cross-sectional study of preprints and final journal publications from COVID-19 studies: discrepancies in results reporting and spin in interpretation. *BMJ Open.* 2021;11(7):e051821.

https://doi.org/10.1017/dmp.2023.82 Published online by Cambridge University Press