Contents lists available at ScienceDirect

EClinicalMedicine

journal homepage: https://www.journals.elsevier.com/eclinicalmedicine

Commentary

# Data explosion during COVID-19: A call for collaboration with the tech industry & data scrutiny

Elizabeth M. Hechenbleikner[a,*], Daniel V. Samarov[b], Ed Lin[a]

[a] Department of Surgery, Division of General and GI Surgery, Emory University School of Medicine, 550 Peachtree Street NE, MOT 9th Floor, Atlanta, GA 30308, United States
[b] Daniel V Samarov, Inc., Atlanta, GA, United States

ARTICLE INFO

*Article History:*
Received 20 April 2020
Revised 29 April 2020
Accepted 30 April 2020
Available online 24 May 2020

The volume and speed of data generation in biomedical literature, social media, and other resources during the COVID-19 pandemic is unprecedented. This mountain of data is growing daily across PubMed, Twitter, Google Scholar, and the World Health Organization's COVID-19 database [1], naming a few. The recently published COVID-19 Twitter dataset may offer insights into multiple topics from compliance with social distancing to assembling homemade masks and mental health tips [2]. Beyond social media, the massive *COVID-19 Open Research Dataset (CORD-19)* has been assembled from tech giants like Microsoft, the Allen Institute for Artificial Intelligence, and Georgetown University's Center for Security and Emerging Technology [3]. This dataset houses over 12,000 full text articles in "machine-readable form" that can be ingested programmatically into computer software programs and analyzed using machine learning applications like natural language processing (NLP). Furthermore, *CovidSurg* is a global registry for tracking outcomes in COVID-19 infected surgical patients with over 100 countries registered [4]. This registry represents a unique opportunity to evaluate variation in patient characteristics, peri-operative management and surgical outcomes. Additionally, guidelines continue to emerge from large international surgical societies like Society of American Gastrointestinal and Endoscopic Surgeons (SAGES). SAGES has developed peri-operative safety practices involving filtration, smoke evacuation, and personal protective equipment use [5]. It is paramount that prospective data collection efforts across these resources and multiple areas of clinical practice continues both institutionally and globally.

Each of these datasets represents a rich repository for application of data mining, machine learning (ML) and artificial intelligence (AI) algorithms. Data scientists and experts in AI and ML are the gate keepers to making this happen. Over the last decade there has been tremendous progress in these areas thanks to advances in computing capabilities and key advances in ML, such as deep learning (DL), a framework for working with complex, unstructured data (to include language/text). DL has pushed the state of the art in NLP, allowing computers to provide insight into complex language patterns, extract topics, understand context and identify relationships of interest. For example, using DL and/or ML methods, one could assess global geographic variation in filtration practices for surgical smoke and associated patient outcomes thus helping improve some of the emerging peri-operative safety guidelines.

While data scientists can help apply technology toward organizing data more efficiently, it is paramount that this information is further filtered by objective individuals and organizations to convey only the facts, the real facts, and remove all the other nonsense that is being portrayed as "fact." Even preeminent organizations that are presumably playing by the books including surgical societies, government organizations, and hospitals are strapped by self-interest, being forced to take action of "facts" while keeping their doors open. While self-interest has some advantages pushing Big Pharma in a race toward developing new vaccines and therapeutic interventions quickly, in parallel, one hopes that data gathered from many sources is being dispersed appropriately and that organizations are acting together to share ideas, technology, and study designs. The reality is that there is insufficient data gathered on COVID-19 and as "facts" continue to emerge each organization is analyzing this information in a decentralized fashion leaving room for varying interpretation and real-life application. Within the United States, government organizations like the Centers for Disease Control and Prevention Division of Healthcare Quality Promotion [6] are capable of vetting the highest quality and most relevant data, helping sort "fact" from fiction, for certain aspects of operational readiness like infection control practices in the peri-operative setting; task forces built around this typically include experts in materials management, hospital infection prevention control, environmental engineering, surgeons, infectious disease clinicians and/or pharmacists. While these larger task forces may be able to provide guidelines, implementation is decentralized at the hospital level. Personnel with similar qualifications should be deployed to ensure guidelines are implemented with attention to local priorities while factoring in guidance from hospital administration and departmental leadership.

In summary, large-scale data collection, organization, scrutiny, and dissemination efforts during the COVID-19 era are key for strategic decision-making and sharing best practices. Due to centralized sources like *CORD-19*, COVID-19 Twitter, and institution-level data, we are

[*] Corresponding author.
*E-mail address:* ehechen@emory.edu (E.M. Hechenbleikner).

https://doi.org/10.1016/j.eclinm.2020.100377

uniquely positioned for collaboration with data scientists and the tech industry. Moreover, separate organizations free from monetary gain, policymaking power, and other personal benefit are further needed to help examine the deluge of data, sorting “fact” from fiction.  While large governmental organizations may have more resources to continually monitor emerging science and data, healthcare organizations are charged with quickly adapting to new guidelines while protecting the health of patients and staff on the front lines.

## Acknowledgements

None of the authors have any relevant conflicts of interest, financial relationships, or funding sources to report for this article.

## Author contributions

All authors contributed equally to this manuscript and meet all four criteria for authorship based on ICMJE Recommendations

## References

[1] World Health Organization. Global research on coronavirus disease (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov. Accessed April 15, 2020.

[2] Chen E., Lerman K., Ferrara E. COVID-19: the first public coronavirus twitter dataset.arXiv.org2020; http://arXiv:2003.07372.

[3] White House Office of Science and Technology Policy. Call to action to the tech community on new machine readable COVID-19 dataset. https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/. Accessed March 31, 2020.

[4] NIHR Global Health Research Unit on Global Surgery. CovidSurg. https://globalsurg.org/covidsurg/. Accessed March 30, 2020.

[5] Society of American Gastrointestinal and Endoscopic Surgeons. Resources for smoke & gas evacuation during open, laparoscopic, and endoscopic procedures. https://www.sages.org/resources-smoke-gas-evacuation-during-open-laparoscopic-endoscopic-procedures/?fbclid=IwAR3K7iqjwvhcqqqUDAq-Tr89u5Enj0CKxmDr3oRG2BlrAP2tOU70JvyYqBwU. Accessed April 1, 2020.

[6] Division of healthcare quality promotion (DHQP). Activities. https://www.cdc.gov/ncezid/dhqp/index.html. Accessed April 27, 2020.