FPF Privacy & Pandemics - Call for Position Statements on Responsible Uses of Technology and Health Data During Times of Crisis

Challenges in Access to Patient-Level COVID-19 Data, and the Role of De-Identification

Niamh McGuinness[1,2] and Luk Arbuckle[1]

[1]Privacy Analytics, 251 Laurier Avenue West, Suite 200, Ottawa ON K1P5J6, Canada
[2]Corresponding Author Email: nmcguinness@privacy-analytics.com

The expeditious and efficient sharing of COVID-19 data will continue to shape the response to the pandemic. The importance of such knowledge exchange was highlighted in past health emergencies, such as the 2014-2015 Ebola epidemic in West Africa.[1] However, despite the global proliferation in coronavirus research, data generation and dissemination have been complicated by several factors. International collaborations, which account for over 20% of all scientific publications[2] have already been impacted by the pandemic, as face-to-face meetings across geopolitical boundaries have been hindered by travel restrictions. Research teams tasked with solving this gargantuan problem are consequently smaller than those who worked on other coronaviruses in the past.[3] Another issue is the complexities that arise when sharing patient-level data across jurisdictions and organizations.

In response to the current COVID-19 pandemic, data protection authorities (DPAs) from around the world have adapted or clarified their requirements.[4] Acknowledging the importance of a vigorous response to the pandemic, some DPAs have responded by pointing out that existing privacy law already has the flexibility to make any necessary adjustments to privacy practices.[5] Regulators around the world are adapting and relaxing their requirements to enable research, and to expedite development and approval of potential new treatments.[6] Notably, the US Department of Health and Human Services announced that, on a discretionary basis, they would not impose their usual penalties for violation of the HIPAA Privacy Rule for "good faith uses and disclosures" of PHI during the pandemic.[7] While this may seem to suggest that de-identification could be ignored, there are many unfortunate examples of harm coming to

[1] WHO, 'Developing Global Norms for Sharing Data and Results during Public Health Emergencies' (*World Health Organization*, 2015) <https://www.who.int/medicines/ebola-treatment/blueprint_phe_data-share-results/en/>.
[2] Dalmeet Singh Chawla, 'International Collaborations Growing Fast' [2018] Nature Index <https://www.natureindex.com/news-blog/international-collaborations-growing-exponentially>.
[3] Virginia Gewin, 'The Trials of Global Research under the Coronavirus' [2020] Nature <https://www.nature.com/articles/d41586-020-02326-0>.
[4] IAPP, 'DPA Guidance on COVID-19' (*International Association of Privacy Professionals*, 2020) <https://iapp.org/resources/article/dpa-guidance-on-covid-19/>; Future of Privacy Forum, 'COVID-19 Privacy & Data Protection Resources' (*Future of Privacy Forum - Privacy and Pandemics*, 2020) 19 <https://sites.google.com/fpf.org/covid-19-privacy-resources/>.
[5] OPC, Canada, 'A Framework for the Government of Canada to Assess Privacy-Impactful Initiatives in Response to COVID-19' (*Office of the Privacy Commissioner of Canada*, April 2020) <https://www.priv.gc.ca/en/privacy-topics/health-genetic-and-other-body-information/health-emergencies/fw_covid/>.
[6] FDA, 'Coronavirus Treatment Acceleration Program (CTAP)' (*U.S. Food and Drug Administration*, 2020) <https://www.fda.gov/drugs/coronavirus-covid-19-drugs/coronavirus-treatment-acceleration-program-ctap>.
[7] HHS, 'OCR Announces Notification of Enforcement Discretion to Allow Uses and Disclosures of Protected Health Information by Business Associates for Public Health and Health Oversight Activities During The COVID-19 Nationwide Public Health Emergency' (*U.S. Department of Health and Human Services*, 2020).

individuals known to be diagnosed with COVID-19, and when the pandemic is over those sharing and releasing data may be held to account for their practices if perceived as overly permissive.

The importance of sharing patient-level COVID-19 data to drive evidence-based decision making has been highlighted by data science consortia and researchers alike.[8] However, both groups agree that data with this level of granularity have not been widely available. While publication of de-identified clinical documents has been mandated under regulations such as European Medicines Agency Policy 0070 and Health Canada Public Release of Clinical Information,[9] structured clinical trial data is predominantly shared on a voluntary basis. Due to the highly personal and potentially identifying nature of such data, secondary use is strictly controlled across jurisdictions. Either explicit consent is sought from the data subjects for such re-use, which may be impractical (or even impossible), or the data are de-identified to the degree that they are classified as non-personal and therefore not subject to the same restrictions.

There are some notable cases where de-identified patient-level COVID-19 data have been successfully shared, such as on secure platforms like Vivli who launched a dedicated portal where such assets are hosted free of charge.[10] It is generally accepted that to derive as much usefulness from de-identified data as possible, a statistical approach, which considers not only the data but the sharing or release context, is the optimal strategy.[11] Understanding the context of data sharing or release is critical for assessing whether vulnerabilities are realistic and could be exploited. The concept of evaluating vulnerabilities and putting them in context is one that is well understood in the field of data security. It is an approach that has also been identified as critical for moving the debate forward with a more meaningful focus on the *process* of de-identification and risk.[12]

Timely de-identification of large clinical datasets such that they can be released quickly to contribute to COVID-19 research efforts is likely to prove a challenge for most data custodians, and this, as well as general concerns about patient privacy, may explain why so few datasets have been made available to date. For example, simple methods such as aggregating counts introduce risks many may not be aware of,[13] and such data may have limited usefulness in practice. The use of privacy models, of which there are many to choose from,[14] becomes even more critical with more detailed and complex data, and empirical

[8] Research Data Alliance, 'RDA COVID-19 Recommendations and Guidelines' (2020) <https://www.rd-alliance.org/system/files/RDA%20COVID-19%3B%20recommendations%20and%20guidelines%2C%205th%20release%20%28final%20draft%29%2028%20May%202020.pdf>; Christopher V Cosgriff, Daniel K Ebner and Leo Anthony Celi, 'Data Sharing in the Era of COVID-19' (2020) 2 The Lancet Digital Health e224.

[9] Health Canada, 'Health Canada Public Release of Clinical Informatiom: Guidance Document' (*Health Canada*, 12 March 2019) <https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance/document.html>; EMA (n 9).

[10] Vivli, 'Share Your COVID-19-Related Trials on the Vivli Portal' (*Vivli*, 26 March 2020) <https://vivli.org/vivli-covid-19-portal-2/>.

[11] Institute of Medicine, *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Consensus Study Report.* (The National Academies Press 2015).

[12] Stephen Bamford, 'Applications of Privacy-Enhancing Technology to Data Sharing at a Global Pharmaceutical Company' (2020) 3 Journal of Data Protection and Privacy 281.

[13] Luk Arbuckle, 'Aggregated Data Provides a False Sense of Security' (*International Association of Privacy Professionals*, 27 April 2020) <https://iapp.org/news/a/aggregated-data-provides-a-false-sense-of-security/>.

[14] Isabel Wagner and David Eckhoff, 'Technical Privacy Metrics: A Systematic Survey' (2018) 51 ACM Computing Surveys.

testing is needed to validate assumptions.[15] New methods such as synthetic data are also emerging that are suitable for some use cases, but again privacy risks exist since data could be replicated or allow for the correct attribution of personal information.[16]

Whereas the level of expertise and knowledge required to deploy de-identification at scale can be daunting for the variety of data needed, the profession is maturing with training and resources.[17] The degree to which the data should be transformed will depend upon the identifiability, which falls on a spectrum.[18] Depending upon the context in which the data will be shared or released, and the security and contractual controls in place, it may be possible to preserve data utility, thus maximizing the benefit of the data for secondary use.

Privacy requirements should not be viewed merely as obstacles to overcome on the journey to sharing useful data. Rather, data custodians should survey the privacy landscape with a lens wide enough to encompass those regulations which have been developed with purpose of protecting the personal rights of data subjects while simultaneously fostering secondary research. While performing statistical de-identification often requires expert application of risk assessment, privacy models, and software, we must also look beyond the technical, and apply privacy best practices to establish trust from the public in the research and medical communities that are working toward finding effective treatments for COVID-19.

[15] Janice Branson and others, ‘Evaluating the Re-Identification Risk of a Clinical Study Report Anonymized under EMA Policy 0070 and Health Canada Regulations’ (2020) 21 Trials <https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-020-4120>.

[16] Nicolas Ruiz, Krishnamurty Muralidhar and Josep Domingo-Ferrer, ‘On the Privacy Guarantees of Synthetic Data: A Reassessment from the Maximum-Knowledge Attacker Perspective’ (Springer 2018).

[17] HITRUST Academy, ‘Data De-Identification Methodology’ (*HITRUST Academy*) <https://hitrustalliance.net/hitrust-academy/de-identification-methodology-course/>.

[18] Future of Privacy Forum, ‘A Visual Guide to Practical Data Da-Identification’ (2017) <https://fpf.org/wp-content/uploads/2017/06/FPF_Visual-Guide-to-Practical-Data-DeID.pdf>.