Journal of the American Medical Informatics Association, 00(0), 2020, 1–4
doi: 10.1093/jamia/ocaa260
Perspective

Perspective

Ten principles for data sharing and commercialization

Curtis L. Cole,[1] Soumitra Sengupta,[2] Sarah Rossetti (née Collins),[2] David K. Vawdrey,[3] Michael Halaas,[4] Thomas M. Maddox,[5] Geoff Gordon,[6] Trushna Dave,[7] Philip R.O. Payne,[8] Andrew E. Williams[9] and Deborah Estrin[10]

[1]Healthcare Policy and Research, Cornell University, New York, New York, USA, [2]Columbia University Irving Medical Center, New York, New York, USA, [3]Geisinger Health System, Danville, Pennsylvania, USA, [4]Stanford University School of Medicine, Palo Alto, California, USA, [5]Healthcare Innovation Lab, BJC HealthCare/Washington University School of Medicine, St. Louis, Missouri, USA, [6]Informatics Institute, University of Alabama-Birmingham, Birmingham, Alabama, USA, [7]IT Business Solutions, NewYork-Presbyterian Hospital, New York, New York, USA, [8]Institute for Informatics (I2), Washington University School of Medicine, St. Louis, Missouri, USA, [9]Tufts Medical Center, Boston, Massachusetts, USA and [10]Cornell Tech, Cornell University, New York, New York, USA

*Corresponding Author: Curtis L. Cole, MD, Weill Cornell Medicine, Wood Library, 1300 York Ave, New York, NY 10065, USA; ccole@med.cornell.edu

Received 7 August 2020; Revised 7 August 2020; Editorial Decision 2 October 2020; Accepted 2 October 2020

ABSTRACT

Digital medical records have enabled us to employ clinical data in many new and innovative ways. However, these advances have brought with them a complex set of demands for healthcare institutions regarding data sharing with topics such as data ownership, the loss of privacy, and the protection of the intellectual property. The lack of clear guidance from government entities often creates conflicting messages about data policy, leaving institutions to develop guidelines themselves. Through discussions with multiple stakeholders at various institutions, we have generated a set of guidelines with 10 key principles to guide the responsible and appropriate use and sharing of clinical data for the purposes of care and discovery. Industry, universities, and healthcare institutions can build upon these guidelines toward creating a responsible, ethical, and practical response to data sharing.

Key words: data sharing, electronic medical records, privacy, policy, intellectual property

INTRODUCTION

The shift from paper to digital medical records has enabled a giant leap in the broad utility of clinical data. The remarkable advances in what is possible with these data has led to increasing demands for healthcare institutions to share data for purposes beyond clinical care, including quality improvement and biomedical research and development. With more data sharing taking place, concerns are being raised about topics such as data ownership, the loss of privacy, and the protection of the intellectual property (IP) that may be encapsulated within or derived from health data. Healthcare institutions are struggling with competing data sharing and privacy demands.

Through discussions at national conferences and invited forums, our group has identified the following common questions related to the sharing of data by healthcare institutions:

- When *must* data be shared for purposes beyond clinical care?
- What rights do providers, patients, and organizations retain when data are compelled to be shared?
- When *can* data be shared for purposes beyond clinical care (ie, education, research, or commercialization)? What obligations do institutions have to their patients and their care providers with respect to privacy, ownership, intellectual property, scientific publication, and transparency?

At the Clinical Research Forum, AMIA, and invited forums hosted by NewYork Presbyterian Hospital, we have discussed these questions with numerous healthcare institutions and discovered considerable variation both between and within organizations in how they answer these questions. We have also observed frustration in regards to the lack of clear guidance from government entities that sometimes create conflicting mandates related to these issues.

One type of conflict arises when there is a requirement for a healthcare institution to share data with a state- or federally-sponsored registry or exchange. Often there is no financial transaction in these agreements and providers may have little leverage over contract terms, absent industry guidelines such as we are proposing. Those responsible for the registry can have relatively unregulated access to use the data therein for purposes that may be inconsistent with our institutional policies. Because institutions are compelled to provide data to specific registries, they have little leverage to constrain how those data are used—and potentially subsequently disclosed—by the operator of the registry.

To help resolve this conundrum, we developed a set of guidelines drawn from multiple institutions and discussed at multiple forums by stakeholders including informaticists, data scientists, clinical researchers, attorneys and other leaders from academic medicine, EHR vendors, regulators, and other related parties. We have aggregated and condensed the knowledge generated into this set of guidelines for the responsible and appropriate use and sharing of clinical data for the purpose of care and discovery—including potential commercialization.

This article outlines these 10 guidelines with some discussion of what is intended and also what is unresolved. It is our hope that industry, universities, and healthcare institutions will adopt these guidelines.

BACKGROUND ASSUMPTIONS

These guidelines are primarily targeted to academic medical institutions that are generally not-for-profit and organized around the shared missions of patient care, education, and research for the purpose of generalizable knowledge. The guidelines embrace existing regulations requiring that data collected in government-funded research be available to others. Similarly, the guidelines honor the principles of patient privacy and ethical research outlined in the Belmont Report.[1]

The goals underpinning the guidelines are: to promote sharing within the principle of "minimum necessary"; to enable discovery without compromising the trust and autonomy of our patients and providers; to foster innovation (including appropriate compensation for innovators); and to preserve opportunities for our investigators to participate in such research.

These goals are difficult. The details are inherently filled with fuzzy boundary conditions. Much like ethical standards in biomedical research, robust discussion, oversight, and self-awareness will be essential to make the spirit of these guidelines come alive in actionable forms, and the application may vary somewhat from instance to instance. These guidelines are a work in progress, but we believe they provide a foundation on which to build.

Mission-driven

1. Data sharing with external parties must be consistent with the organization's core missions of patient care, education, and research for the purpose of generalizable knowledge and the advancement of health.

While each academic medical center (AMC) is unique, most are not-for-profit institutions with a similar tripartite mission of clinical care, research, and education. While financial sustainability is important, AMCs are typically not incorporated to generate profit for shareholders. In light of their not-for-profit status, activities that are not clearly linked to AMCs' core missions must be considered carefully. Any activity that requires sharing clinical data should be consonant with at least 1 of the institutional missions. If patient data are shared, the Belmont Report principle of "beneficence" applies, obligating the institution to "maximize possible benefits."[1]

Payment for academic work

2. Financial compensation should be based on the value of the contribution (eg, academic research, expertise, or invention) provided by the mission-driven organization; data alone and financial gain should not be primary drivers.

AMCs routinely receive funding to conduct research, such as clinical trials of medications; these activities are consonant with the mission of an AMC. Trials are expensive and it is appropriate that they are funded by the corporations that may subsequently financially benefit from drug and device sales. Moreover, clinical trials provide patients with access to experimental therapies that may fundamentally change the course of their disease. AMCs provide the staff, infrastructure, and data required to perform these studies.

AMCs can help advance cutting-edge research in fields, such as AI and Machine Learning, adding value to data analytics and algorithm development. Departments of informatics, data science, information technology, biostatistics, and others have local knowledge and unique insight regarding how datasets have been created and therefore how to optimally use or leverage our unique clinical data assets. Not-for-profit institutions make significant investments to collect and curate this data, continuously investing in maintaining and advancing these assets so they may be used for discovery.

It is appropriate to recover the cost of providing the data and expertise externally or for invention, but institutions should not go so far as to merely sell data for profit. We believe that patients may individually opt to do this on their own, but that should be their decision, not ours.

Minimum necessary

3. Data sharing will be limited to the minimum data elements needed for the project.

As with HIPAA, the principle of "minimum necessary" should be applied to data sharing agreements. Pragmatic attempts to deidentify data should be used whenever possible. This concept should allow for practical interpretation.

Limited agreements

4. Data sharing agreements should be nonexclusive, have defined time limits, and permission for data use should be revocable at any time.

The conditions surrounding data sharing can change over time. A recipient may change ownership or mission. Social standards may evolve, especially as technology advances and unimagined uses of data become possible over time. As such, data sharing agreements should not be perpetual and should be carefully scoped to limit use to a specific purpose. They should contain fixed expirations, with

appropriate data destruction provisions and ongoing oversight of data usage through audits. Agreements can be renewed after reexamination determines that the context remains appropriate.

No transfer of ownership

5. Data sharing agreements confer stewardship; data ownership cannot be transferred and, as such, recipients cannot redistribute or sell the data.

The notion of "ownership" of clinical data is fraught with complexity. For example, a physician's interpretation of a pathology slide or diagnostic image might be perceived as ownership by the physician, the entities who paid him or her, entities involved in the creation of the image, and/or the patient involved. A wise practice might be to refer to the "stewardship of data" rather than outright ownership. Patients are free to use their data as they like, but they cannot completely control how an AMC uses the same data; it is information authored and collected as part of the institution's core business and is captured and documented into the proprietary infrastructure and information systems of the AMC. Likewise, it is not appropriate for healthcare institutions to transfer ownership of that which is not uniquely theirs. Therefore, recipients of data must know that they do not own the data, and they have no rights to redistribute or sell it, including as part of derivative data sets that are created through analysis or other use of the data.

No reidentification

6. Data recipients should not attempt to reidentify deidentified data.

One of the critical lessons of the last few years is that deidentification is relative and does not constitute true anonymization.[2,3] As a practical matter, it may take significant effort to reidentify data. But more and more entities are capable of doing just that, particularly when leveraging existing public data sets outside of healthcare. Therefore, agreements must contractually prohibit reidentification and include the potential for audit and penalties applied for misuse.

Limited data association

7. Data cannot be associated with other data sets without explicit permission.

The association of various data sets is an extremely powerful tool for discovery, but is fraught with potential for abuse. Depending on the types of data that are linked, there is potential to yield insights about an individual outside the scope of predetermined research or services agreements or extend beyond the reasonable mission-related activities. Therefore, data sharing agreements must explicitly define what will be allowed and anything beyond that must be prohibited.

Transparency

8. The key purpose of data sharing activities and engagements should be transparent to all stakeholders, including patients and study participants.

Recent revelations that some providers have been sharing data without transparency[4] have understandably triggered the outrage of patients and may undermine the trust of our entire industry. AMCs are incorporated for the public good. Few legitimate care or research activities consistent with our missions and public trust should happen in secret. While the details of patentable inventions and other intellectual property might not be immediately made public, the fact that patient data is being used for discovery or invention must never be hidden. The basic intentions and structure of work involving patient data should be visible to all stakeholders, including patients themselves.

A more difficult issue involves certain types of work, such as algorithm development. The very nature of some algorithms can make them inherently difficult, if not impossible, to understand.[5] Poorly designed algorithms have the potential to exacerbate inherent bias and inequality.[6] Special effort and oversight are required to ensure that factors that affect how the algorithms were created and are used are transparent to those affected.

Conflicts

9. Conflicts of interest must be transparent with appropriate governance of both employee and organization-level conflicts.

As with pharmaceutical use and testing, data analysis and use can involve financial interests or other forms of conflict that might affect objectivity or skew intentions away from the best interests of patients or the mission of an organization. While oversight alone is not sufficient to prevent these effects, transparency regarding conflicts is a minimum requirement with some mechanism for managing these conflicts, be they at the individual or institutional level.

Oversight

10. All decisions about data sharing should be overseen by appropriate representative stakeholders – much like an institutional review board overseeing human subjects research.

Guidelines such as those presented here are inherently imperfect. Not only do they fail to address every relevant question, but real-world scenarios routinely have ambiguity and present tensions between different guidelines. Therefore, as with human subjects research, we recommend that institutions establish a data sharing review committee. Similar to an Institutional Review Board, a data sharing review committee should be composed of qualified individuals, with appropriate expertise and representation, to oversee data sharing requests that challenge internal guidelines and policy. The committee should continuously optimize compliance and evolve with inherent changes in the field, while staying true to the spirit implied in the guidelines. It would be wise to include patient representation in the committee.

CONCLUSION

These 10 principles are not exhaustive. If healthcare providers and universities can build upon this initial set of principles—as a group—then we have a chance of holding accountable our own faculty, staff, and leaders in addition to the myriad business partners with whom we desire or are obligated to share data.

FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

All the authors participated in the conception, drafting, or revising of the content and have approved the version to be published. The opinions expressed are designed to provoke a conversation working toward a consensus within the field, but do not reflect the opinions of the authors' individual institutions or employers.

ACKNOWLEDGMENTS

The authors wish to thank Andrea Messina, MBA, and Kimberly Durniak, PhD, of Mass General Brigham for sharing their insights.

CONFLICT OF INTEREST STATEMENT

Dr. Maddox discloses he is currently employed as a cardiologist and the executive director of the Healthcare Innovation Lab at BJC HealthCare/Washington University School of Medicine. In this capacity, he is advising Myia Labs, for which his employer is receiving equity compensation in the company. He is receiving no individual compensation from the company. He is also a compensated director for a New Mexico-based foundation, the JF Maddox Foundation.

Dr. Estrin discloses that she is a part-time employee of Amazon.

REFERENCES

1. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Department of Health, Education, and Welfare; 1979. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html Accessed October 23, 2020.
2. Janmey V, Elkin PL. Re-identification risk in HIPAA de-identified datasets: The MVA attack. *AMIA Annu Symp Proc* 2018; 2018: 1329–37.
3. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. *J Med Internet Res* 2019; 21 (5): e13484.
4. Copeland R. Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans. *Wall Street Journal* 2019; https://www.wsj.com/articles/google-s-secret-project-nightingale-gathers-personal-health-data-on-millions-of-americans-11573496790? mod=hp_lead_pos1 Accessed October 23, 2020.
5. Diakopoulos N, Koliska M. Algorithmic transparency in the news media. *Digital Journalism* 2017; 5 (7): 809–28.
6. Matheny MS, Thadaney Israni S, Ahmed M, Whicher D, eds. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. Washington, DC: National Academy of Medicine; 2019.