Commentary

mHealth Systems Need a Privacy-by-Design Approach: Commentary on "Federated Machine Learning, Privacy-Enhancing Technologies, and Data Protection Laws in Medical Research: Scoping Review"

Ambuj Tewari[1,2], PhD

[1]Department of Statistics, University of Michigan, Ann Arbor, MI, United States

[2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Ambuj Tewari, PhD
Department of Statistics
University of Michigan
1085 S University Ave
Ann Arbor, MI, 48109-1107
United States
Phone: 1 734 615 0928
Email: tewaria@umich.edu

Related Article:
Comment on: http://www.jmir.org/2023/1/e41588/

Abstract

Brauneck and colleagues have combined technical and legal perspectives in their timely and valuable paper "Federated Machine Learning, Privacy-Enhancing Technologies, and Data Protection Laws in Medical Research: Scoping Review." Researchers who design mobile health (mHealth) systems must adopt the same privacy-by-design approach that privacy regulations (eg, General Data Protection Regulation) do. In order to do this successfully, we will have to overcome implementation challenges in privacy-enhancing technologies such as differential privacy. We will also have to pay close attention to emerging technologies such as private synthetic data generation.

(J Med Internet Res 2023;25:e46700) doi: 10.2196/46700

KEYWORDS

mHealth; differential privacy; private synthetic data; federated learning; data protection regulation; data protection by design; privacy protection; General Data Protection Regulation; GDPR compliance; privacy-preserving technologies; secure multiparty computation; multiparty computation; machine learning; privacy

Introduction

Brauneck et al [1] should be congratulated for reviewing privacy-enhancing technologies (PETs) from a legal standpoint. The right to privacy is a fundamental human right, the importance of which in the current digital age cannot be overstated. Protecting this basic human right will need the cooperation of scholars and experts from many disciplines. It is therefore heartening to see legal experts joining hands with technical experts to engage in a thoughtful discussion of how General Data Protection Regulation (GDPR) legislation in the European Union relates to commonly used PETs including federated learning (FL), differential privacy (DP), and secure multiparty computation (SMPC).

The GDPR recognizes that privacy should be a primary design consideration when designing systems that deal with personal data. Privacy is not something to be added on as an afterthought once the system has already been designed. Researchers in the health sciences, especially mobile health (mHealth), are beginning to adopt a "privacy-by-design" mindset. My own group at the University of Michigan and my clinical collaborators have started to seriously study privacy in the context of mHealth [2,3], but much remains to be done.

Differential Privacy

Brauneck et al [1] correctly point out that FL alone does not sufficiently protect user privacy. This is well known. In fact, the original paper that proposed FL itself pointed out that FL

will have to be supplemented with technologies such as DP and SMPC to achieve adequate privacy protection. In this commentary, I will primarily focus on DP. Since I am not a legal expert, my comments will necessarily be from a technical perspective.

DP and its variants have emerged as a leading PET. It has been adopted by technology companies such as Apple and Google. The US Census Bureau also chose it for the 2020 US Census. Calls to revisit foundational statistical theory to incorporate privacy constraints have also formulated the problem using DP [4].

DP has some clear strengths. It is a clear formalism with desirable theoretical properties and increasing software support. However, the epsilon parameter in DP is hard to interpret in the context of specific applications. Its mathematical meaning is precise, but it is often very hard to choose a good value of epsilon to achieve a careful balance between privacy and statistical utility. Researchers have proposed building an "Epsilon Registry" to help the community make sensible implementation choices [5]. More community efforts, especially from the medical informatics community, will be needed to successfully realize the potential of DP.

It is also important to note that recent DP literature is nicely complemented by older statistics literature on statistical disclosure control [6]. It is unlikely that a one-size-fits-all solution will emerge for all data protection scenarios. It is therefore important for system designers to have a broad understanding of available tools. Moreover, old and new tools need to be examined from a legal perspective just as Brauneck et al [1] have done for FL, DP, and SMPC. This is challenging because technology and the law are both undergoing changes. Hopefully, PETs and privacy laws will coevolve so that society will benefit from the ongoing data revolution without threats to the fundamental human right to privacy.

Private Synthetic Data

Brauneck et al [1] do not mention private synthetic data generation as a PET, but I believe that private synthetic data has tremendous potential for enabling data-driven innovation in health care without sacrificing privacy. The use case the authors considered is one where a data processing workflow (eg, FL) needs to be modified to ensure that it satisfies DP. A different use case is where we simply publish synthetic data that "is similar" to the original sensitive data but which protects user privacy (eg, in the DP sense). This way, downstream data analysts do not have to modify their workflows and can simply work with the synthetic data just as they would with the original data.

However, what does it mean to "be similar" to the original data set? One possibility is that one might hope to preserve correlations between attributes. For a while, it was thought that this could only be done using methods that will be computationally intractable. However, there is recent progress in this area [7], which has renewed interest in the possibility of generating statistically useful synthetic data that nevertheless provably protects user privacy.

Conflicts of Interest

None declared.

References

1. Brauneck A, Schmalhorst L, Majdabadi MMK, Bakhtiari M, Völker U, Baumbach J, et al. Federated Machine Learning, Privacy-Enhancing Technologies, and Data Protection Laws in Medical Research: Scoping Review. J Med Internet Res 2023;25:e41588 [FREE Full text] [doi: 10.2196/41588]
2. Liu JC, Goetz J, Sen S, Tewari A. Learning from others without sacrificing privacy: simulation comparing centralized and federated machine learning on mobile health data. JMIR Mhealth Uhealth 2021 Mar 30;9(3):e23728 [FREE Full text] [doi: 10.2196/23728] [Medline: 33783362]
3. Shen A, Francisco L, Sen S, Tewari A. Exploring the relationship between privacy and utility in mobile health: a simulation of federated learning, differential privacy, and external attacks. J Med Internet Res (forthcoming) 2023. [doi: 10.2196/43664]
4. Wainwright MJ. Constrained forms of statistical minimax: computation, communication and privacy. In: Proceedings of the International Congress of Mathematicians. 2014 Presented at: ICM 2014; Aug 13-21, 2014; Seoul, South Korea URL: https://people.eecs.berkeley.edu/~wainwrig/Barcelona14/Wainwright_ICM14.pdf
5. Dwork C, Kohli N, Mulligan D. Differential privacy in practice: expose your epsilons. JPC 2019 Oct 20;9(2). [doi: 10.29012/jpc.689]
6. Slavković A, Seeman J. Statistical data privacy: a song of privacy and utility. Annu Rev Stat Appl 2022 Nov 18;10(1). [doi: 10.1146/annurev-statistics-033121-112921]
7. He Y, Vershynin R, Zhu Y. Algorithmically effective differentially private synthetic data. arXiv. Preprint posted online Feb 11, 2023. [doi: 10.48550/arXiv.2302.05552]

Abbreviations

DP: differential privacy
FL: federated learning
GDPR: General Data Protection Regulation
mHealth: mobile health

PET: privacy-enhancing technology
SMPC: secure multiparty computation

Edited by T Leung; this is a non–peer-reviewed article. Submitted 21.02.23; accepted 22.02.23; published 30.03.23.

Please cite as:
Tewari A
mHealth Systems Need a Privacy-by-Design Approach: Commentary on "Federated Machine Learning, Privacy-Enhancing Technologies, and Data Protection Laws in Medical Research: Scoping Review"
J Med Internet Res 2023;25:e46700
URL: https://www.jmir.org/2023/1/e46700
doi: 10.2196/46700
PMID: