A source data privacy framework for synthetic clinical trial data

Afrah Shafquat[1], Jason Mezey[1,2,3], Mandis Beigi[1], Jimeng Sun[1,4], Jacob Aptekar[1]

[1] Medidata Solutions, a Dassault Systèmes company
[2] Cornell University, Ithaca, NY
[3] Weill Cornell Medicine, New York, NY
[4] University of Illinois Urbana Champaign, Champaign, IL

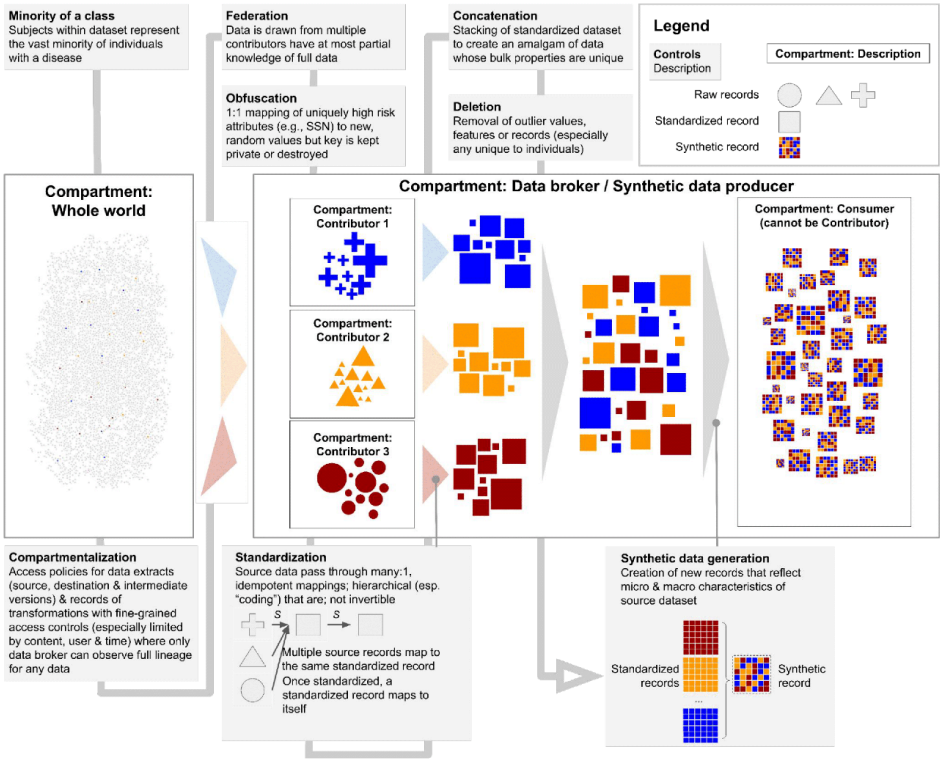
Background:

Synthetic data generation, the production of realistic data from a real data source, has shown promise in boosting performance of machine learning models by improving the quality[1,2,3] and quantity of training samples while preserving the privacy of the individuals. In the context of clinical trials, the ability to share synthetic clinical trial data while preserving patient privacy has been touted as a strategy for improving drug safety, evaluating bias, and other meta-analysis of multiple clinical trial studies[4,5]. Distinct from other industries, the presence of protected health information requires a conservative approach when sharing data, even synthetic data, in the healthcare domain. To ensure the privacy of clinical trial data, an open and effective privacy standard is needed such that it (i) protects patients from unwanted disclosures and financial or personal harm, (ii) allows institutions to contribute data by upholding their legal and ethical commitments to their patients, and (iii) supports the adoption of realistic synthetic data as an effective means of sharing useful information while protecting critical privacy interests (e.g., identities of clinical trial sponsors, clinical trial studies, and patients). In this paper, we propose a source data privacy framework that increases the privacy protection upstream of synthetic data generation at the level of the source data. Consequently, the overall privacy of the generated synthetic dataset is inherited from the privacy delivered through this framework, enhancing privacy mechanisms intrinsic to the synthetic data generation model.

Privacy System:

The proposed privacy framework aims at increasing privacy using a series of technical, policy, and algorithmic controls upstream of the synthetic data generation process such that the source data ingested by the synthetic data generator protects the privacy of the contributor-level and study-level attributes even in its raw, unaltered form. The transformations are targeted toward improving the resulting dataset privacy and utility. We assume that Electronic Data Capture (EDC) data is available to data contributors as case report forms (CRFs), where data contributors (and their respective customers) may only have access to CRFs for their own trials or those of their partners. Apart from a data leak within the data broker (entity responsible for synthetic data generation), the attacker may never access the data generated from the privacy framework components. Figure A summarizes the proposed privacy framework where the components are **(i) Minority of class, (ii) Compartmentalization, (iii) Federation, (iv) Obfuscation, (v) Standardization, (vi) Concatenation, (vii) Deletion, and (viii) Synthetic data generation.**

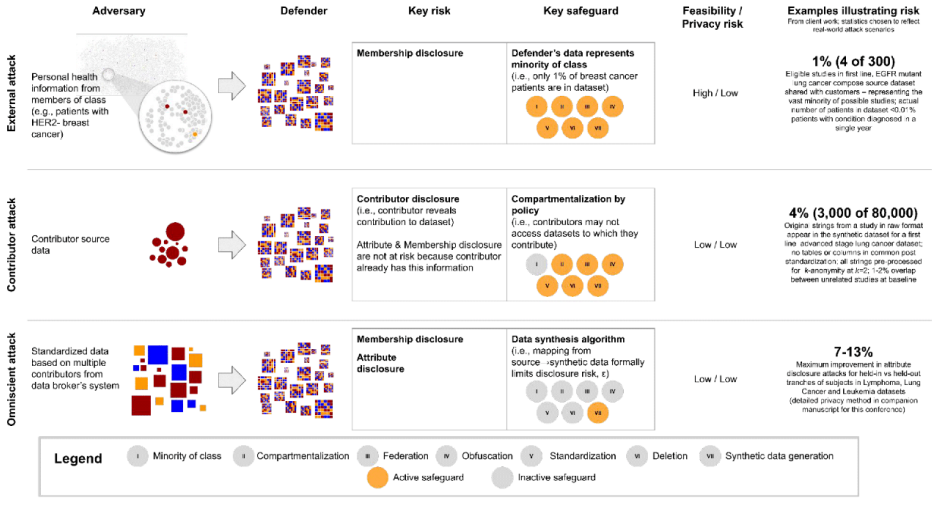
Overview of privacy system design



Attack Scenarios:

Considering 77% of all data breaches in 2015-2019 were in the healthcare sector [7] and the continued increase in the number of data breaches and costs associated [8-10], we consider all likely attack scenarios, attack adversaries, and the corresponding key disclosure risks. We summarize possible attacks as: (i) External attack, (ii) Contributor attack, and (iii) Omniscient attack. Figure B summarizes these adversarial scenarios where disclosure risk in each category (i.e., Membership, Contributor, and Attribute) is described. The examples demonstrate how the proposed privacy system exerts robust control over the most damaging and likely attack scenarios.

Attack scenarios



Conclusion:

Sharing clinical trial data presents a unique challenge for preserving multiple layers of privacy i.e., on the individual-level, study-level, and contributor-level. Although the framework can accommodate other types of data, the proposed privacy system is specifically designed to address the challenges in clinical trial data synthesis.

References

1. G.D., et al. PLoS One 2022
2. ET et al. CoRR 2021
3. T. J. et al AAAI 2021
4. H.H.W et al HHS. 2014
5. Z.A. et al BMJ Open 2021
6. OCR 2018
7. A.H.S. et al. Healthcare 2020
8. J.M. Health IT Security 2022
9. J.M. Health IT Security 2022
10. J.M. Health IT Security 2022