View in browser

Global Alliance for Genomics & Health

GDPR Brief

Federated analysis for responsible data sharing under the GDPR

April 2022



Melissa Cline



The General Data Protection Regulation (GDPR) presents a number of restrictions on how organizations both within and outside of the European Union (E.U.) may process (i.e. collect, use and share) personal data, which is defined as data that relates to "an identified or identifiable person". While these restrictions present obstacles to sharing genomic and health data, federated analysis can offer a solution. Traditional data sharing involves data providers sending a copy of their data to data recipients, who analyze the data at their home institutions ("bringing the data to the code"). Federated analysis, conversely, involves "bringing the code to the data", with requesting parties submitting a copy of their analysis software to the data, and with the data not shared beyond the host institution. Federated analysis shares the aggregate, group-level results of data analysis amongst collaborating institutions, without revealing the individual-level personal data used to perform this analysis. Therefore, federated data analysis enables research institutions to engage in collaborative data analysis without engaging in the exchange of personal biomedical data, which may facilitate GDPR compliance. For instance, this could in some instances reduce the number of participants in data analysis which the law considers to be joint data controllers.

Two examples are the Beacon network and the Matchmaker Exchange, in which a number of different organizations host services that allow specific queries of their data, enabling the discovery of cases presenting a rare variant or symptoms suggesting a rare disease.

Often, there is one institution that oversees data coordination by organizing and issuing data requests and collating and disseminating the results. This approach is used in the CanDIG and CINECA networks. Within these networks, software containers are shared with the partner organizations, each of which apply them to their internal data within their secure institutional environments, generating anonymized data or contributing to its de-identification. The analysis results can then be shared across the network, and are also harmonized to common technical standards. The GA4GH encourages federation for the sharing of data that "cannot move for technical or legal reasons".

For aggregated data to not be regulated as personal data, there must no available means to infer the identities of the underlying individuals in the group from the aggregate results, that is reasonably likely to be used. This is not necessarily true for aggregate data about rare diseases or rare genetic variants, which might plausibly be observed in just one person. If data are organized according to demographic traits such as ethnicity or age bracket, it might be possible to infer the identities of the concerned individuals from a unique combination of demographic traits belonging to them. As such, aggregated data are not always anonymized. As with personal data, the privacy of aggregated data or data that have been de-identified are best evaluated with a contextual risk-based approach. Within data science, the field of "statistical disclosure control" (SDC) offers an expansive literature and a breadth of methods for reducing the risk of disclosing personal data in data sharing, in balance with ensuring that the data to be shared remains informative.

Using federated analysis methods instead of disclosing identifiable personal data also provides other advantages in the alignment of research priorities and GDPR compliance. For example, using federated data analysis methodologies that limit the processing of personal data can facilitate compliance with the Data Minimization principle established in Article 5 of GDPR.

A well-designed federated analysis strategy that leverages open-source software can promote the safety and integrity of the research process through in enhancing the reproducibility and transparency of the output results.

In summary, while federated analysis alone does not guarantee compliance with data protection law, a well-designed federated analysis mechanism can enable responsible data sharing that is in compliance with the GDPR.

Melissa Cline is a Program Manager at the University of California Santa Cruz.

Please note that GDPR Briefs neither constitute nor should be relied upon as legal advice. Briefs represent a consensus position among Forum Members regarding the current understanding of the GDPR and its implications for genomic and health-related research. As such, they are no substitute for legal advice from a licensed practitioner in your jurisdiction.

Further Reading

- Emanuela Podda. "Shedding light on the legal approach to aggregate data under the GDPR & the FFDR. UNECE Conference on European Statistics (2021).
- Mike Hintze. "Viewing the GDPR through a De-Identification Lens: A Tool for Compliance, Clarification, and Consistency" International Data Protection Law (2018).
- Giovanni Comandè and Giulia Schneider. "Differential Data Protection Regimes in Data-Driven Research: Why the GDPR Is More Research-Friendly Than You Think" German Law Journal (2021).

Relevant GDPR Provisions

- Article 4(1) - Definition of personal data
- Article 5 - Principles relating to processing of personal data
- Recital 162 - Processing for Statistical Purposes

Past Briefs

What constitutes an 'international data transfer?'

February 2022



Processing the Data of the Deceased

December 2021



For a list of all previous briefs, please consult here.

Global Alliance for Genomics and Health, 661 University Avenue, Suite 510, Toronto, ON M5G 0A3, Canada

Unsubscribe Manage preferences