Enabling Open Science in Medicine Through Data Sharing: An Overview and Assessment of Common Approaches from the European Perspective

Hammam Abu Attieh[1°], Anna Haber[1°], Felix Nikolaus Wirth[1°], Benedikt Buchner[2], Fabian Prasser[1*]

[1]*Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Health Data Science Center, Medical Informatics Group, Charitéplatz 1, 10117 Berlin, Germany*
[2]*University of Augsburg, Chair for Civil Law, Liability Law and Law of Digitization, Universitätsstraße 2, 86159 Augsburg, Germany*

[*]*Corresponding author. E-Mail: fabian.prasser@bih-charite.de*
[°]*Contributed equally to this work.*

Abstract. Open Science involves the sharing of knowledge and data as well as the exchange of research results. This is particularly important in the biomedical field, as it can foster validation studies in response to the replication crisis and improve resource utilisation. Since medical data is particularly privacy sensitive, its processing is subject to strong data protection requirements. Agencies, institutions, and projects in the European Union are still struggling with the establishment of widely accepted mechanisms supporting the sharing of data for Open Science practices. The goal of this paper is to provide an overview of different methods that have been used for this purpose and to discuss their technical properties and legal challenges. Our assessment is based on well-known conceptualizations, such as the Five Safes Framework. The result shows that different approaches provide different trade-offs between the functionalities and the degree of data protection provided, and that there are open legal issues. Current legislative initiatives in the EU, including regulations for the European Health Data Space and the Data Governance Act, have the potential to address some of the resulting uncertainties.

Keywords: Open Science, Medicine, Data Sharing, Law and Technology

1. Introduction

1.1 Introduction

Open Science refers to the use of transparent and collaborative methods of producing, using, and communicating data as well as results and it is seen as one of the major mechanisms to end the replication crisis (Coiera et al.). Open Science is especially important in data-driven biomedical research, a field that relies heavily on the collaborative and shared use of data beyond the purpose for which it was originally collected, such as the use of healthcare data for generating scientific evidence (Gewin; Merson et al.; Taichman et al.). Some researchers see data sharing as an ethical necessity (Bauchner et al.) and argue that it saves lifes (Besançon et al.). Moreover, it is promoted by various institutions and funding agencies (Hulsen; Institute of Medicine).

However, medical data is particularly sensitive and in the European Union (EU) the processing and sharing of personal data requires a legal basis in compliance with the requirements laid out in the EU General Data Protection Regulation (GDPR). The most common mechanism for this is obtaining an informed consent by the data subject (see Articles 6, 7 and 9 GDPR). However, obtaining consent is often challenging and

costly, e.g., when data is to be analysed retrospectively (Williams and Pigeot), and patient-centric solutions, e.g., for obtaining dynamic consent through apps or portals have proven to be difficult to implement in practice (Teare et al.). In addition, the legitimation basis of consent is also problematic because consent is only effective if it is given voluntarily, in an informed manner and for a specific purpose, but in practice it is often not possible to ensure that these conditions are met.

While the GDPR defines some further grounds permitting data processing, also for the use of data for research purposes (see Art. 89 and Recital 159 GDPR), these are interpreted differently by the Member States or can be regulated country-specifically due to opening clauses (Baker and Mckenzie). As a result of this heterogeneity, it is very difficult to share data broadly based on such legal bases. Country-specific special laws (e.g., for national cancer registries) only apply to some types of data and their application is usually tied to additional restrictions.

Due to these legal challenges, current large-scale data sharing efforts, are usually based on technological approaches that aim to provide anonymity for the data subjects (Wirth et al., 2021). As a consequence, no personal data is shared and the process is not subject to the requirements of the GDPR and other national laws. Moreover, such solutions adhere to the principle of data minimisation, according to which the processing of non-anonymized data is only lawful to the extent that the specific (scientific) purpose of the processing cannot be equally achieved with anonymized data.

1.2 Background

Providing anonymity is challenging. In recent years, it has increasingly been accepted that anonymity cannot (solely) be a data property (Ohm; Rubinstein and Hartzog). This means that, for complex datasets, anonymity of subjects can only be provided when going beyond measures that apply directly to the data (such as anonymization techniques) by carefully designing and considering the process in which the data is used and its contextual embedding. On the technical side, this is reflected in the *differential privacy* concept, which is a mathematical property of (randomized) data processing algorithms. If correctly parameterized, it ensures that the output of a processing algorithm doesn't disclose sensitive information about individual data subjects (Dwork).

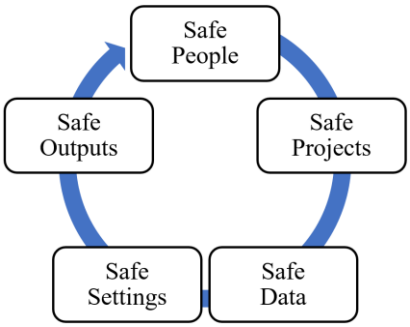

Figure 1. Elements of the five safes framework.

However, process-oriented measures for providing anonymity can also be designed along further technical and also organisational axes. A

common way to formalize this is the Five Safes Framework, which has been developed for reasoning about privacy protection when providing access to data for research purposes (Desai et al.). As is illustrated in Figure 1, it specifies five dimensions of protection, the so-called "Safes":

1. **Safe people:** Only trusted and qualified researcher should be provided access to sensitive data.
2. **Safe projects:** Data access should only be provided for appropriate, legitimate, and ethical purposes.
3. **Safe data:** Shared data should be protected to the highest degree possible on the data level, e.g., through anonymization or pseudonymisation.
4. **Safe settings:** Data is shared in special "settings" that provide additional protection from unauthorized access or disclosure of sensitive information.
5. **Safe outputs:** Scientific outputs derived from shared data, such as statistics published in scientific papers, should not disclose sensitive information.

The Five Safes Framework is a valuable tool for reasoning about the degree of anonymity provided while processing sensitive data. According to Recital 26 of the GDPR, "all means reasonably likely" should be considered based on "objective factors" when determining whether data sets can be related to an identifiable person. We note that the term "safe" as used by the framework is to be understood as an umbrella term for providing trustworthiness, security and privacy protection. We further note that there are some legal uncertainties and heterogeneous interpretations regarding the importance of individual axes of the framework. For example, it remains controversial to which degree the motivation of anticipated adversaries (cf. "Safe People") can be taken into account (Spindler and Schmechel). Moreover, different actors have different roles (data controller, data processor, joint data controller) and different responsibilities that need to be clearly defined. In this context, technical solutions that ensure privacy can also serve as a means to mitigate responsibilities.

Along the dimensions outlined above, various (technical) methods have been developed for sharing data while preserving the anonymity data subjects. These methods have different advantages and disadvantages, are associated with further legal challenges, and none can solve the problem as a "silver bullet". In the following section, we will present some of the most common approaches and highlight their legal assumptions and open challenges.

2. Overview and assessment

2.1 Anonymization

A traditionally common approach to sharing data anonymously is to focus on the data level (cf. "Safe Data") and modify the data itself to such a degree that it can be considered non-personal. This process is illustrated in Figure 2 and it can involve techniques such as the deletion, aggregation, generalisation, or perturbation of values and variables (Fung et al.).

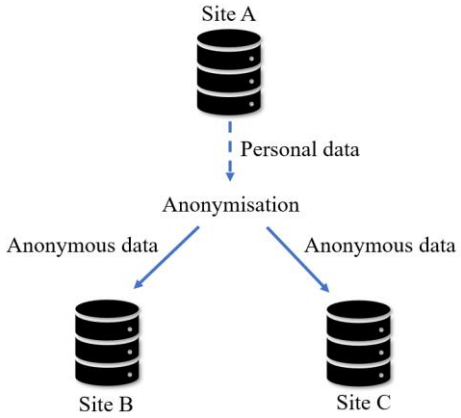

Figure 2. Schematic illustration of data sharing through anonymization. In this example, site A shares an anonymised dataset with sites B and C.

However, anonymization requires trading off the degree of protection achieved against the usefulness of the data to be shared (e.g., in terms of the preservation of its statistical properties) (Fung et al.). The more comprehensive and complex a dataset is, the more challenging it becomes to achieve a reasonable trade-off (Ohm; Rubinstein and Hartzog) and it can be expected that only focusing on the data level doesn't work for many sharing scenarios in the biomedical domain (Wirth et al., 2021).

A recent application of anonymization for Open Science in medicine includes public use files provided by the LEOSS registry (Jakob et al.) and the German National Pandemic Cohort Network (NAPKON) (Koll et al.) on COVID-19 patients. The datasets were anonymised following a mixed methodology combining quantitative and qualitative approaches using ARX, a relatively well-known software for the anonymization of structured medical datasets (Prasser et al.). It must be noted, however, that in line with the limitations described above, the anonymous datasets created in these projects are relatively narrow and only contain a low number of variables.

On the legal side, some general open issues regarding anonymity and anonymization have already been outlined in the previous section. In addition, it is challenging to align technical anonymization approaches on the data level with legal concepts around the identifiability of data.

As a further example, the Opinion on Anonymization Techniques by the Article-29 Working Party outlines requirement regarding the protection of data, including protection from "singling out", "linkability" and "inference" (Art. 29 WP). Formalizing these requirements can be challenging and remains an open problem (see work by (Cohen and Nissim) for an example regarding the concept of "singling out"). More concrete specifications, guidelines and certification bodies would be needed to provide researchers aiming to share data with reliable answers to these questions. According to the European Data Protection Board, anonymisation of personal data might be difficult to achieve and maintain "due also to ongoing advancements in available technological means, and progress made in the field of re-identification." Anonymisation of personal data should therefore be approached "with caution" (EDPB).

2.2 Distributed data analyses

Distributed or federated data networks are mechanisms focusing on "Safe Outputs" which are specifically suitable for creating shared data pools across multiple participating institutions. Each node of such a network maintains its own dataset without direct access for the other parties.

However, as is illustrated in Figure 3, methods are supported for calculating joint statistical results on the overall dataset by transferring specifically designed analysis scripts or commands to the participating institutions and merging the individual (already aggregated) results to obtain a global result (e.g., deriving a global mean for a certain variable from the local means of the participating nodes).

Figure 3. Schematic illustration of data sharing with a distributed data analysis infrastructure. In the example, site A receives a joint result for the data from sites B and C.

Examples of distributed solutions that have been developed for sharing data in the biomedical field include DataSHIELD (Gaye et al.), the tools by the Observational Health Data Sciences and Informatics (OHDSI) initiative (Hripcsak et al.) and the Personal Health Train (PHT) (Beyan et al.).

Distributed data networks are used on a broad scale to share data and generate biomedical knowledge (see, for example, (Burn et al.; Hong et al.; Oluwagbemigun et al.)). However, they also suffer from necessary trade-offs between their usefulness and the degree of anonymity provided and from open legal issues. Most importantly, also statistical analysis results can potentially disclose sensitive information about data subjects (e.g., statistical tables with low cell counts (Smith et al.)). Thus, although the platforms exchange only aggregated data between nodes, it is possible that the data still qualifies as personal data under the GDPR (cf. Art. 4 (5) GDPR). The platforms therefore must implement restrictions regarding the types of analyses that can be performed and specific protection mechanisms that ensure that only "Safe Outputs" are shared (cf. Recital 162 GDPR).This means, that not all types of medically relevant statistical analyses can be performed using such platforms and that they are subject to legal uncertainties regarding the degree of anonymity provided analogously to approaches that are based on data-

level anonymization.

2.3 Cryptographic approaches

More recently, cryptographic protocols have been developed that enable joint calculations on encrypted data hence focusing on the "Safe Setting" aspect (Canetti). Examples include Secure Multi-Party Computation (SMPC) methods and Homomorphic Encryption (HE). As is illustrated in Figure 4, conceptually, the institutions aiming to create common (virtual) data pool, encrypt their data and execute specific algorithms, potentially requiring multiple rounds of communication.

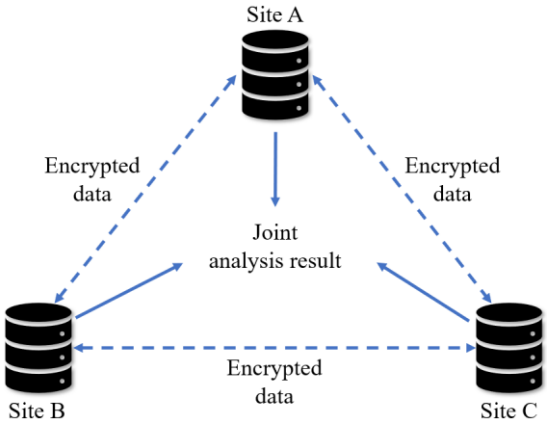

Figure 4. Schematic illustration of data sharing with cryptographic methods. In this example, sites A, B and C calculate a common statistical result.

Examples of cryptographic data sharing platforms include MedCo (Raisaro et al.) which supports a broad range of medical and bioinformatics analyses and EasySMPC (Wirth et al., 2022), which provides fewer functionalities but is more easy to install and operate.

Compared with the other approaches, also cryptographic solutions suffer from inherent trade-offs between protection and usefulness. In this specific case, however, this is mostly because cryptographic platforms are complex and can be difficult to operate and extend with additional functionalities. If implemented correctly, the data sharing process itself can be mathematically proven to be secure and not disclose any data. Whether the joint statistical result also maintains the subjects' anonymity depends on the specific types of analysis supported, which is important to ensure when operating such platforms. Helminger and Rechberger have proposed a test in form of a flow chart (Helminger and Rechberger) to handle this issue.

On the legal side, the question remains open to whether the join processing of encrypted data can be considered as anonymous processing. Several authors have brought forward according arguments for cryptographic data sharing mechanisms (1) in general (Spindler and Schmechel; Helminger and Rechberger), (2) in the biomedical domain (Scheibner et al.) (3) or in other areas (Treiber et al.).

2.4 Data enclaves

Data enclaves are another data sharing approach relying specifically on the creation of a "Safe Setting". With this approach, as is illustrated in

Figure 5, data is basically pooled centrally in its original or pseudonymised form in a secure and trusted environment. Researchers can then apply for data access and are provided with secure access mechanisms, e.g., including monitored remote desktop connections. It is important to point out, that the enclave stores personal data while the researcher will only be provided with access to aggregated, non-identifiable data. Data enclaves are often operated by public institutions or official agencies.

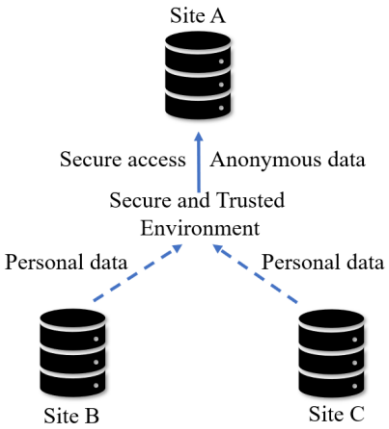

Figure 5. Schematic illustration of sharing data through an enclave. In the example, sites B and C share their data with site A.

Well-known examples of Safe Havens for biomedical data include the Scottish National Safe Haven (Public Health Scotland), operated by the Scottish National Health Services, or the US Center for Medicare and Medicaid Service Virtual Research Data Center (*CMS Virtual Research Data Center*), operated by the US Center for Medicare and Medicaid Service. In Germany, the enclave concept is implemented by the Center for Cancer Registry Data and the Research Data Center of the Federal Office for Drug Research and Medical Devices (BfArM) (BfArM; *Cancer Registry Data*).

In regards to the provided functionalities data enclaves are more capable than the approaches presented previously. This is due to the fact that they can process personal individual-level data, which, for example, enables linkage across datasets. The situation is more complicated on the risk-side of the trade-off, however, which is reflected in legal challenges.

From the legal perspective, it remains unclear whether the processing of (pseudonymized) individual-level data in an enclave that provides secure access with further safeguards ensuring that researchers cannot identify individuals can be considered “anonymous processing”. This would require a truly process-oriented assessment of the degree of identifiability and the option to put a strong focus on the security of the enclave and the trustworthiness of the institution operating it. Moreover, classification as "anonymous processing" can only be considered if one follows the so-called relative approach, according to which the perspective *of the individual data processor* must be taken into account in order to decide whether data are personally identifiable or not (in contrast to the so-called absolute approach, according to which data must always be classified as personally identifiable if *anyone in the world* can identify a data subject

in the data set (Helminger and Rechberger)).

As a consequence of this legal uncertainty, enclaves are usually only operated today based on special laws that provide according permissions. In Germany, for example, the operation of the Center for Cancer Registry Data is based legally on the Cancer Registry Act and the Research Data Center for Health is based legally on specific paragraphs in the German Social Code (SGB).

3. Discussion

The legislative efforts of the EU and its Member States show a clear intention to foster the utilization of data for research purposes. At the same time, improvements in privacy-enhancing technologies provide increasingly more favourable trade-offs between privacy risks and the usefulness of data sharing platforms. These are good signs for a broader adoption of Open Science principles in medicine.

However, current legislation is too abstract and fragmented (Paseri; Aurucci) in some areas and has led to heterogeneous interpretations in the EU Member States due to opening clauses. In order to further expand data sharing and Open Science in the medical field, harmonization efforts, guidelines and certification bodies are needed. An important recent development is the proposal for the European Health Data Space (EHDS) which builds upon the GDPR, the Data Governance Act, the draft EU Data Act, and the NIS Directive (see Proposal Regulation COM/2022/197 final). The EHDS proposal builds upon many of the concepts outlined in this article, such as the sharing of data through Secure Processing Environments (SPEs) (see Art. 50 Proposal Regulation COM/2022/197 final). The legal framework for sharing data laid out in the draft EU Data Act and the EHDS proposal have the potential to overcome some of the legal issues around technical solutions available.

Acknowledgements

This work has partly been done as part of the NFDI4Health Consortium (www.nfdi4health.de). We gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — project number 442326535. This work has also been partly funded by the European Union's Horizon 2020 research and innovation programme under the project ORCHESTRA — grant agreement No 101016167.

References

Art. 29 WP (Article 29 Working Party). "Opinion 05/2014 on Anonymisation Techniques." *European Commission* (2014).

Association of Population Based Cancer Registries in Germany. *Common ADT-GEKID Record Format*. Web. Accessed 20 Jan. 2023. https://www.gekid.de/common-adt-gekid-record-format.

Aurucci, Paola. "Legal issues in regulating observational studies: The impact of the GDPR on Italian biomedical research." *Eur. Data Prot. L. Rev*. 5 (2019): 197.

Becker, Regina, et al. "Applying GDPR roles and responsibilities to scientific data sharing." *International Data Privacy Law* 12.3 (2022): 207-219.

Besançon, Lonni, et al. "Open science saves lives: lessons from the COVID-19

pandemic." *BMC Medical Research Methodology* 21.1 (2021): 1-18.
Bauchner, Howard, Robert M. Golub, and Phil B. Fontanarosa. "Data Sharing: An Ethical and Scientific Imperative." *Journal of the Medical Association* 315.12 (2016): 1238-1240, doi:10.1001/jama.2016.2420.
Beyan, Oya, et al. "Distributed Analytics on Sensitive Medical Data: The Personal Health Train." *Data Intelligence* 2.1-2 (2020): 96-107, doi: https://doi.org/10.1162/dint_a_00032.
BfArM (Federal Institute for Drugs and Medical Devices). *Health Data Lab*. Web. Accessed 28 Jan. 2023. https://www.healthdatalab.de/.
Burn, Edward, et al. "Opioid Use, Postoperative Complications, and Implant Survival after Unicompartmental Versus Total Knee Replacement: A Population-based Network Study." *The Lancet Rheumatology* 1.4 (2019): e229-e236.
Cancer Registry Data. Web. Accessed 28 Jan. 2023. https://www.krebsdaten.de/Krebs/EN/Home/homepage_node.html.
Canetti, Ran. "Security and Composition of Multiparty Cryptographic Protocols." *Journal of Cryptology* 13 (2000): 143-202.
CMS Virtual Research Data Center. Web. Accessed 28 Jan. 2023. https://www.resdac.org/cms-virtual-research-data-center-vrdc.
Cohen, Aloni, and Kobbi Nissim. "Towards Formalizing the Gdpr's Notion of Singling Out." *Proceedings of the National Academy of Sciences* 117.15 (2020): 8344-8352.
Coiera, Enrico, et al. "Does Health Informatics Have a Replication Crisis?". *Journal of the American Medical Informatics Association* 25.8 (2018): 963-968.
Desai, Tanvi, Felix Ritchie, and Richard Welpton. "Five Safes: Designing Data Access for Research." *Economics Working Paper Series* 1601 (2016): 28.
Dwork, Cynthia. "Differential Privacy: A Survey of Results." *TAMC'08: Proceedings of the 5th international conference on Theory and applications of models of computation*, Xi'an, China, April 25 - 29, 2008. Edited by Manindra Agrawal, et al., Springer, 2008, pp 1-19.
EDPB (European Data Protection Board). "Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research." EDPB Documents 2 February 2021.
Fung, Benjamin CM, et al. "Introduction to Privacy-preserving Data Publishing: Concepts and Techniques." *Chapman and Hall/CRC*, 2010.
Gaye, Amadou, et al. "DataSHIELD: Taking the Analysis to the Data, Not the Data to the Analysis." *International Journal of Epidemiology* 43.6 (2014): 1929-1944.
Gewin, Virginia. "Data Sharing: An Open Mind on Open Data." *Nature* 529.7584 (2016): 117-119.
Helminger, Lukas, and Christian Rechberger. "Multi-party Computation in the GDPR." *Privacy Symposium 2022: Data Protection Law International Convergence and Compliance with Innovative Technologies (DPLICIT). Cham: Springer International Publishing*, 2022.
Hong, Na, et al. "Preliminary Exploration of Survival Analysis Using the OHDSI Common Data Model: A Case Study of Intrahepatic Cholangiocarcinoma." *BMC Medical Informatics and Decision Making* 18.5 (2018): 81-88.
Hripcsak, George, et al. "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers." *MEDINFO 2015: eHealth-enabled Health. IOS Press*, 2015. 574-578.
Hulsen, Tim. "Sharing Is Caring - Data Sharing Initiatives in Healthcare." *International Journal of Environmental Research and Public Health* 17.9 (2020): 3046.
Downey, Autumn S., and Steve Olson, eds. "Sharing Clinical Research Data: Workshop Summary." *National Academies Press*, 2013.
Jakob, Carolin EM, et al. "Design and Evaluation of a Data Anonymization Pipeline to Promote Open Science On Covid-19." *Scientific Data* 7.1 (2020): 435.
Koll, Carolin EM, et al. "Statistical Biases Due to Anonymization Evaluated in an Open Clinical Dataset from Covid-19 Patients." *Scientific Data* 9.1 (2022): 776.
McKenzie, Baker. "GDPR National Legislation Survey," (2018).".
Merson, Laura, Oumar Gaye, and Philippe J. Guerin. "Avoiding Data Dumpsters—toward Equitable and Useful Data Sharing." *New England Journal of Medicine* 374.25 (2016): 2414-2415.
Ohm, Paul. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA Law Review*. 57 (2009): 1701.

Oluwagbemigun, Kolade, et al. "Dietary Patterns Are Associated with Serum Metabolite Patterns and Their Association Is Influenced by Gut Bacteria among Older German Adults." *The Journal of Nutrition* 150.1 (2020): 149-158.

Paseri, Ludovica. "COVID-19 Pandemic and GDPR: When Scientific Research becomes a Component of Public Deliberation." *Data Protection and Privacy, Volume 14: Enforcing Rights in a Changing World*. Vol. 14. Bloomsbury Publishing (2021): 157-185.

Prasser, Fabian, et al. "Flexible Data Anonymization Using Arx - Current Status and Challenges Ahead." *Software: Practice and Experience* 50.7 (2020): 1277-1304.

Public Health Scotland. *Use of the National Safe Haven.* Web. Accessed 20 Jan. 2023. https://www.isdscotland.org/Products-and-Services/EDRIS/Use-of-the-National-Safe-Haven/.

Raisaro, Jean Louis, et al. "MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16.4 (2018): 1328-1341.

Rubinstein, Ira S., and Woodrow Hartzog. "Anonymization and Risk." *Washington Law Review* 91 (2016): 703.

Scheibner, James, et al. "Revolutionizing Medical Data Sharing Using Advanced Privacy-enhancing Technologies: Technical, Legal, and Ethical Synthesis." *Journal of medical Internet research* 23.2 (2021): e25120.

Smith, Duncan, and Mark Elliot. "A Measure of Disclosure Risk for Tables of Counts." *Transactions on data privacy* 1 (2008): 34-52.

Spindler, Gerald, and Philipp Schmechel. "Personal Data and Encryption in the European General Data Protection Regulation." *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 7 (2016): 163.

Taichman, Darren B., et al. "Sharing Clinical Trial Data: A Proposal from the International Committee of Medical Journal Editors." *Annals of Internal Medicine* 164.7 (2016): 505-506.

Teare, Harriet JA, Megan Prictor, and Jane Kaye. "Reflections on Dynamic Consent in Biomedical Research: The Story so Far." *European Journal of Human Genetics* 29.4 (2021): 649-656.

Treiber, Amos, et al. "Data Protection Law and Multi-Party Computation: Applications to Information Exchange between Law Enforcement Agencies." *WPES'22: Proceedings of the 21st Workshop on Privacy in the Electronic Society*, Los Angeles, USA, November 7, 2022. Edited by Yuan Hong and Lingyu Wang, Association for Computing Machinery, 2022, pp 69–82.

Williams, Garrath, and Iris Pigeot. "Consent and Confidentiality in the Light of Recent Demands for Data Sharing."*Biometrical Journal* 59.2 (2017): 240-250.

Wirth, Felix Nikolaus, et al. "EasySMPC: A Simple but Powerful No-code Tool for Practical Secure Multiparty Computation." *BMC Bioinformatics* 23.1 (2022): 531.

Wirth, Felix Nikolaus, et al. "Privacy-preserving Data Sharing Infrastructures for Medical Research: Systematization and Comparison." *BMC Medical Informatics and Decision Making* 21.1 (2021): 1-13.