2 Toward Technically Enforceable Consent in Healthcare Research

Johannes Lohmöller, Jan Pennekamp und Klaus Wehrle[1]

2.1 Motivation

Digital health technology has, over the past years, become an integral part of healthcare research, for instance, to gain insight into clinical decision-making (Wang et al., 2022) and its impact on patient trajectories or for large-scale studies, e.g., in cardiovascular research (Denaxas & Morley, 2015). The increasing digitization of healthcare facilities has led to a broad availability of electronic health record (EHR) data and a recent political and public surge to utilize this data. The COVID-19 pandemic further accelerated this trend (Dron et al., 2022). Ethics committees are starting to acknowledge the potential for so-called *secondary use* in medical research, hinting at a broad utilization of EHR data in the (near) future. Currently, there are ongoing state-level initiatives, e.g., in Germany, to centralize EHR data for research purposes (Rau et al., 2024).

Since protecting private medical data is a crucial personal right, a fundamental principle in medical research is giving consent, i.e., individual data subjects agree to a specific use of their data, similar to how they would agree to participate in a clinical trial. For secondary data use, however, the individuals subject to the data can hardly give explicit consent, e.g., as specific research questions have yet to be formulated at the time of data collection. Likewise, individuals often are not reachable anymore once research questions have been fixed for data collected in the past. Here, state-of-the-art practices are general consent or broad consent forms, although these are criticized for being too vague and not specific enough (Barazzetti et al., 2020). Generally, data subjects are willing to consent to secondary use of their data, as highlighted by a recent meta-review (Baines et al., 2024). However, the key to such consent is that individuals retain control over their data and that benefits are clear to them. We thus argue that technical means to enforce consent in healthcare research are highly beneficial. To this end, we investigate such technical means to implement consent in healthcare research that render central data collection and collecting broad consent in advance, as currently discussed (Rau et al., 2024), redundant.

Our work complements distributed data analysis tools, including MedCo (Froelicher, 2020; Raisaro et al., 2019), UnLynx (Froelicher et al., 2017), PCORnet (Yuan et al., 2017), or PHT (Mou et al., 2023), showing the feasibility and need for privacy-preserving decentralized analysis of health records and medical research data. None of them, however, incorporates consent on a research project or even query level. Likewise, related data ecosystems fail to reliably and transparently provide (technical) guarantees to manage patients' consent decisions (Geisler et al., 2022; Lohmöller et al., 2024). In this work, we thus sketch a system design for consent-aware distributed data analysis. Our goal is to provide technical enforcement of consent on a per-query basis while still allowing for large-scale studies. Thereby, we aim to empower users to reliably, transparently, and privacy preservingly handle their consent affinity.

2.2 Sketching Technical Consent Enforcement

Figure 1 introduces the high-level protocol flow, as described in the following. We consider data-holding institutions (e.g., clinics) and data subjects (e.g., patients) each (part-time) active parties in the system such that researchers can only access the data if both parties agree. Thereby, data

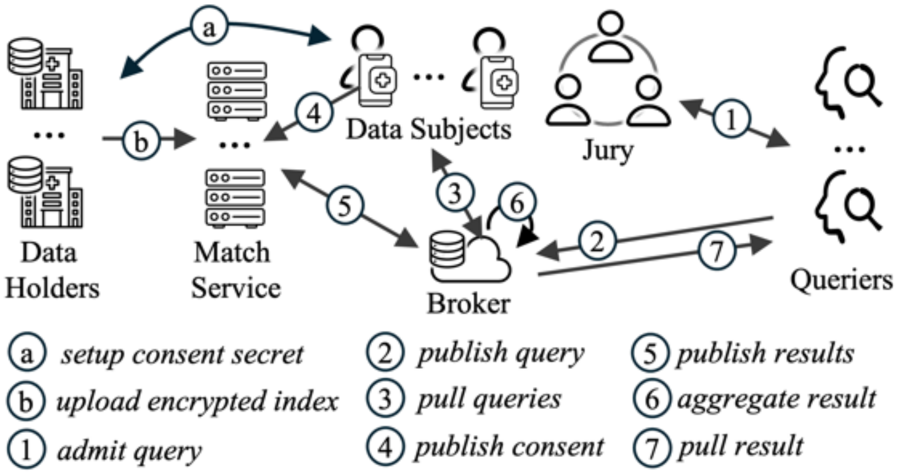
[1] RWTH Aachen University, Aachen, Germany, {lohmoeller,pennekamp,wehrle}@comsys.rwth-aachen.de

subjects remain in control over their data and give consent on a per-query basis. To enforce consent cryptographically, we utilize a public-key searchable encryption scheme (PEKS) that allows defining trapdoors to be matched against an encrypted distributed index (Boneh et al., 2004; Froelicher et al., 2017).

We propose to store the encrypted index under the control of data providers (e.g., hospitals), let researchers submit queries (translated to trapdoors) to the data providers, and cryptographically involve the data subjects in the trapdoor evaluation process, which prohibits query evaluation on data without the data subject's consent. Specifically, we include data subjects in the process by distributing a share of the evaluation key material to the data subject out of band, such as when the data is initially collected. This key share is cryptographically required to evaluate trapdoors, which implies that the data subject needs to actively contribute her share, i.e., consent to query evaluation.

As data subjects can hardly assess whether a specific query is in their interest or beneficial to them, we require an independent jury to review the queries and a short, to a non-expert audience, understandable summary of the research goals. Based on this summary, data subjects can make an informed decision about contributing their data. We employ threshold cryptography to ensure that the jury must agree on a joint query admission decision before a query can be executed. This admission ensures that queries are not directly harmful, such as being exploited for tracing specific individuals. Additionally, we aggregate results from multiple data providers before relaying them to the researchers via the broker shown in Figure 1. This aggregation unlinks the data from any individual and their physical origin, thus preserving the data subjects' anonymity.

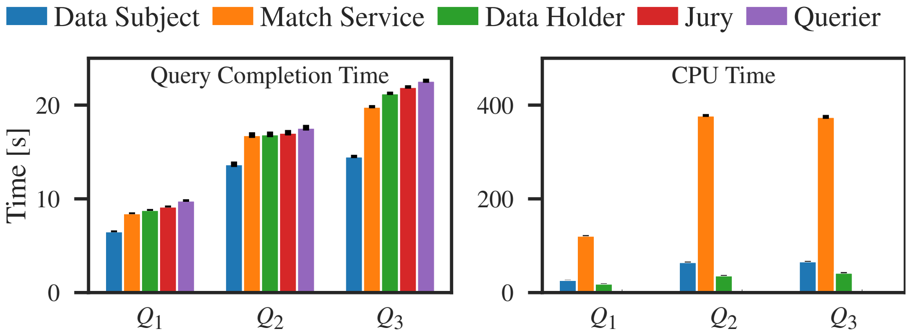
Figure 1: Schematic protocol flow enabling technically enforced consent.



Queries and data are encrypted with a jury-managed public-private key pair. Steps a-b are one-time operations, Steps 1-7 execute per query. Non-collusion is required between data holders and the test servers. Besides, one jury member must be benign.

We evaluate our approach in a simulation study, i.e., we simulate a state-level network composed of German university hospitals. With the HCUP national inpatient sample (Khera et al., 2017), we employ a large dataset of ICD-10 coded diagnoses and evaluate the same real-world queries within the setting as related work before (Yuan et al., 2017). The (one-time) setup for generating keying material and encrypting individual data items consumes 290 ms per record in the dataset, which we consider feasible for practical deployment on a large scale. More interestingly, Figure 2 touches upon the query results of our simulated real-world evaluation, showing that in an idealized scenario (all entities are online and respond immediately), our three queries Q_1(Chung et al., 2015), Q_2 (Habermann et al., 2014), Q_3 (George et al., 2014), which, e.g., analyze the clinical treatment and outcome of a cancer subtype, all execute within 22s. In this setting, the most time-consuming part is matching trapdoors against the encrypted index, followed by the giving consent operation of the data subjects. However, the latter will be spread among up to 6.3 million data subjects in the network, thus incurring a negligible overhead per individual. Practically, we expect the query runtimes to be dominated by user interaction, e.g., after notification via push messaging or regular digests of incoming requests. A smartphone app holding the data subject's keys would be a convenient way of interacting with data subjects. Other options, including delegating consent management to a general physician or relatives, are technically possible and can complement these efforts. Overall, these results show that the suggested approach scales to real-world scenarios.

Figure 2: Query performance for three realistic queries on 6.3 million records.



Left: time passed between retrieval and completion. Right: consumed CPU time (sum over all parallel computations) . By subsampling our dataset, we find that the computationally intensive operations scale linearly with the number of records in the dataset.

2.3 Discussion and the Road Ahead

A frequent concern is that the self-determination and freedom of giving consent, as well as the duty to decide, might lead to consent fatigue. Then, users are overwhelmed by the number of consent requests, similar to Cookie consent banners (Kretschmer et al., 2021). One way to mitigate this issue would be to deploy a local consent agent, e.g., as part of the aforementioned consent management app, that accepts queries based on user-defined rules and criteria the jury has reviewed. Compared to the broad consent practice, data subjects would still be free to review and reconsider their choices (Matzutt et al., 2017). Besides, such an agent would keep the subject's sovereignty high while allowing transparency and accountability.

Compared to a fully centralized system, our proposal at first seems to limit the query experience: While the formulated trapdoors support querying specific ranges and logical formulas for combining multiple matching criteria, a distributed approach that requires human interaction can hardly compete with the performance and interactiveness of a central database. However, efforts to implement centralized EHR databases at the state level or beyond suffer from various issues, including ethical and legal concerns, among others (Baines et al., 2024). Here, data governance strategies are scarce: They need to (1) comply with regulatory frameworks, (2) be perceived as trustworthy by the public, and (3) must not become an overly complicated access barrier for researchers (Rau et al., 2024). We argue that collecting consent on a per-query basis shifts the burden of finding a universal governance strategy to multiple individual choices with clear consequences and limited scope, thereby reducing decision and governance complexity. With this work, we thus sketch a technical basis for this paradigm shift and call for future work to study the data subject's perception of privacy and the willingness to provide their data in more detail.

Besides analyzing implementing the consent agent and studying individual data subjects' perception of privacy, future work should investigate the system's usability with real users, e.g., as part of a clinical trial, and assess its impact on the patient's overall willingness to participate in secondary data use. From a technical perspective, we are confident that the system can be deployed in a real-world scenario, as the evaluation shows good scalability regarding participating entities and data volume.

2.4 Conclusion

Our work complements existing decentralized data analysis tools with enforceable consent, as users are cryptographically involved in the query evaluation. Thereby, it enables data subjects to decide participation freely and on a per-query basis while still allowing for large-scale studies. Thus, our work empowers users to reliably and transparently handle their consent affinity. Our evaluation shows that the computational overhead is reasonable and that the system scales to the demands of real-world scenarios. This work thus contributes a cryptographic option for giving consent to the ongoing discussion on how to open up healthcare data for research (Baines et al., 2024).

Acknowledgments

This research was funded within the VeSiTRUST project and supported under grant no. 02J24A030 by funds of the German Federal Ministry of Education and Research (BMBF). The authors thank Michael Herwig for his initial work on distributed searchable encryption.

Bibliography

Baines, R., Stevens, S., Austin, D., Anil, K., Bradwell, H., Cooper, L., Maramba, I. D., Chatterjee, A., & Leigh, S. (2024). Patient and Public Willingness to Share Personal Health Data for Third-Party or Secondary Uses: Systematic Review. Journal of Medical Internet Research, 26, e50421. https://doi.org/10.2196/50421

Barazzetti, G., Bosisio, F., Koutaissoff, D., & Spencer, B. (2020). Broad consent in practice: Lessons learned from a hospital-based biobank for prospective research on genomic and medical data. European Journal of Human Genetics, 28(7), 915–924. https://doi.org/10.1038/s41431-020-0585-0

Boneh, D., Di Crescenzo, G., Ostrovsky, R., & Persiano, G. (2004). Public Key Encryption with Keyword Search. In C. Cachin & J. L. Camenisch (Eds.), Advances in Cryptology—EUROCRYPT 2004 (pp. 506–522). Springer. https://doi.org/10.1007/978-3-540-24676-3_30

Chung, T. K., Rosenthal, E. L., Magnuson, J. S., & Carroll, W. R. (2015). Transoral robotic surgery for oropharyngeal and tongue cancer in the U nited S tates. The Laryngoscope, 125(1), 140–145. https://doi.org/10.1002/lary.24870

Denaxas, S. C., & Morley, K. I. (2015). Big biomedical data and cardiovascular disease research: Opportunities and challenges. European Heart Journal - Quality of Care and Clinical Outcomes, 1(1), 9–16. https://doi.org/10.1093/ehjqcco/qcv005

Dron, L., Kalatharan, V., Gupta, A., Haggstrom, J., Zariffa, N., Morris, A. D., Arora, P., & Park, J. (2022). Data capture and sharing in the COVID-19 pandemic: A cause for concern. The Lancet Digital Health, 4(10), e748–e756. https://doi.org/10.1016/S2589-7500(22)00147-9

Froelicher, D. (2020). MedCo2: Privacy-Preserving Cohort Exploration and Analysis. Digital Personalized Health and Medicine. https://doi.org/10.3233/SHTI200174

Froelicher, D., Egger, P., Sousa, J. S., Raisaro, J. L., Huang, Z., Mouchet, C., Ford, B., & Hubaux, J.-P. (2017). UnLynx: A Decentralized System for Privacy-Conscious Data Sharing. Proceedings on Privacy Enhancing Technologies. https://petsymposium.org/popets/2017/popets-2017-0047.php

Geisler, S., Vidal, M.-E., Cappiello, C., Lóscio, B. F., Gal, A., Jarke, M., Lenzerini, M., Missier, P., Otto, B., Paja, E., Pernici, B., & Rehof, J. (2022). Knowledge-Driven Data Ecosystems Toward Data

Transparency. Journal of Data and Information Quality, 14(1), 1–12. https://doi.org/10.1145/3467022

George, E. M., Tergas, A. I., Ananth, C. V., Burke, W. M., Lewin, S. N., Prendergast, E., Neugut, A. I., Hershman, D. L., & Wright, J. D. (2014). Safety and Tolerance of Radical Hysterectomy for Cervical Cancer in the Elderly. Gynecologic Oncology, 134(1), 36–41. https://doi.org/10.1016/j.ygyno.2014.04.010

Habermann, E. B., Thomsen, K. M., Hieken, T. J., & Boughey, J. C. (2014). Impact of Availability of Immediate Breast Reconstruction on Bilateral Mastectomy Rates for Breast Cancer across the United States: Data from the Nationwide Inpatient Sample. Annals of Surgical Oncology, 21(10), 3290–3296. https://doi.org/10.1245/s10434-014-3924-y

Khera, R., Angraal, S., Couch, T., Welsh, J. W., Nallamothu, B. K., Girotra, S., Chan, P. S., & Krumholz, H. M. (2017). Adherence to Methodological Standards in Research Using the National Inpatient Sample. JAMA, 318(20), 2011. https://doi.org/10.1001/jama.2017.17653

Kretschmer, M., Pennekamp, J., & Wehrle, K. (2021). Cookie Banners and Privacy Policies: Measuring the Impact of the GDPR on the Web. ACM Transactions on the Web, 15(4), 1–42. https://doi.org/10.1145/3466722

Lohmöller, J., Pennekamp, J., Matzutt, R., Schneider, C. V., Vlad, E., Trautwein, C., & Wehrle, K. (2024). The unresolved need for dependable guarantees on security, sovereignty, and trust in data ecosystems. Data & Knowledge Engineering, 102301. https://doi.org/10.1016/j.datak.2024.102301

Matzutt, R., Müllmann, D., Zeissig, E.-M., Horst, C., Kasugai, K., Lidynia, S., Wieninger, S., Ziegeldorf, J. H., Gudergan, G., gen. Döhmann, I. S., Wehrle, K., & Ziefle, M. (2017). myneData: Towards a Trusted and User-controlled Ecosystem for Sharing Personal Data. https://doi.org/10.18420/IN2017_109

Mou, Y., Li, F., Weber, S., Haneef, S., Meine, H., Caldeira, L., Jaberansary, M., Welten, S., Yediel Ucer, Y., Prause, G., Decker, S., Beyan, O., & Kirsten, T. (2023). Distributed Privacy-Preserving Data Analysis in NFDI4Health With the Personal Health Train. Proceedings of the Conference on Research Data Infrastructure, 1. https://doi.org/10.52825/cordi.v1i.282

Raisaro, J. L., Troncoso-Pastoriza, J. R., Misbach, M., Sousa, J. S., Pradervand, S., Missiaglia, E., Michielin, O., Ford, B., & Hubaux, J.-P. (2019). MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16(4), 1328–1341. https://doi.org/10.1109/TCBB.2018.2854776

Rau, E., Tischendorf, T., & Mitzscherlich, B. (2024). Implementation of the electronic health record in the German healthcare system: An assessment of the current status and future development perspectives considering the potentials of health data utilisation by representatives of different stakeholder groups. Frontiers in Health Services, 4, 1370759. https://doi.org/10.3389/frhs.2024.1370759

Wang, J., Yang, L., Huang, X., & Li, J. (2022). Annotating Free-Texts in EHRs Towards a Reusable and Machine-Actionable Health Data Resource. In P. Otero, P. Scott, S. Z. Martin, & E. Huesing (Eds.), Studies in Health Technology and Informatics. IOS Press. https://doi.org/10.3233/SHTI220239

Yuan, J., Malin, B., Modave, F., Guo, Y., Hogan, W. R., Shenkman, E., & Bian, J. (2017). Towards a privacy preserving cohort discovery framework for clinical research networks. Journal of Biomedical Informatics, 66, 42–51. https://doi.org/10.1016/j.jbi.2016.12.008