German Medical Data Sciences 2022 - Future Medicine
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI220802

Secure Multi-Party Computation Based Distributed Feasibility Queries – A HiGHmed Use Case

Reto WETTSTEIN[a,1,2] , Tobias KUSSEL[b,2], Hauke HUND[c], Christian FEGELER[c], Martin DUGAS[a], and Kay HAMACHER[b]
[a] *Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany*
[b] *Computational Biology & Simulation, Technical University Darmstadt, Darmstadt, Germany*
[c] *GECKO Institute, Heilbronn University of Applied Sciences, Heilbronn, Germany*

Abstract. The integration of routine medical care data into research endeavors promises great value. However, access to this extra-domain data is constrained by numerous technical and legal requirements. The German Medical Informatics Initiative (MII) – initiated by the Federal Ministry of Research and Education (BMBF) – is making progress in setting up Medical Data Integration Centers to consolidate data stored in clinical primary information systems. Unfortunately, for many research questions cross-organizational data sources are required, as one organization's data is insufficient, especially in rare disease research. A first step, for research projects exploring possible multi-centric study designs, is to perform a feasibility query, i.e., a cohort size calculation transcending organizational boundaries. Existing solutions for this problem, like the previously introduced feasibility process for the MII's HiGHmed consortium, perform well for most use cases. However, there exist use cases where neither centralized data repositories, nor Trusted Third Parties are acceptable for data aggregation. Based on open standards, such as BPMN 2.0 and HL7 FHIR R4, as well as the cryptographic techniques of secure Multi-Party Computation, we introduce a fully automated, decentral feasibility query process without any central component or Trusted Third Party. The open source implementation of the proposed solution is intended as a plugin process to the HiGHmed Data Sharing Framework. The process's concept and underlying algorithms can also be used independently.

Keywords. Feasibility queries, distributed processes, privacy, secure multi-party computation, medical informatics, BPMN, FHIR

[1] Corresponding Author, Reto Wettstein, Institute of Medical Informatics, Heidelberg University Hospital, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany; E-mail: reto.wettstein@med.uni-heidelberg.de

[2] These authors contributed equally

1. Introduction

1.1. Background

The era of big data promises vast advancements in nearly all research fields, be it chronic disease management [1], personalized medicine [2], psychiatry [3], or intensive care research [4]. One way to incorporate this paradigm into medical research is to unlock the use of routine medical care data for research purposes [5]. For this reason, the Medical Informatics Initiative (MII) [6] was established by the German Federal Ministry of Education and Research, aiming to connect Germany's university hospitals with research institutes and health-care businesses. The initiative's primary goal is the development of suitable infrastructures and processes to meet the paradigm. The involved university hospitals are establishing so-called Medical Data Integration Centers (MeDICs), in which data from primary medical information systems' are integrated into research repositories using open standards as well as harmonized interfaces and processes [7,8].

The MII's infrastructure tries to aid medical researchers in many steps of the research process. This work is especially concerned with the step of feasibility queries, a preparatory step for clinical studies in order to determine the size of an available cohort. Many research projects require cohort sizes only achievable by consolidating data of multiple organizations. Unfortunately, even if no identifying patient data are processed, disclosure of aggregated data can still become a privacy risk. Especially for studies dealing with rare diseases, the geographical data of organizations can be used for re-identification, due to the very small number of patients treated at each hospital. Hence, a distributed, privacy-preserving feasibility process based on the cryptographic techniques of secure Multi-Party Computation (MPC) is designed, implemented, and tested.

1.2. Objective and Requirements

The MII Taskforce for Process Modeling, on behalf of the National Steering Committee (NSG), developed a high-level process template describing feasibility queries [9]. Additionally, the MII Data Protection Concept (DSK) [10] describes the legal requirements and gives concrete recommendations. Based on these two documents, a fully decentralized process for feasibility queries in small cohort sizes was developed to meet the more specific requirements of the MII's HiGHmed [11] consortium. These requirements are:

1. An automated, fully decentral process should be employed.
2. Patients' sensitive data must be protected with highest privacy guarantees.
3. The privacy of very small (local) cohort sizes should be protected, no Trusted Third Party (TTP) must be used.
4. The process must be deployable on the HiGHmed framework for data sharing.
5. Interoperability should be ensured by using open standards and data models.

To meet these requirements, the process presented in this work was designed and implemented as a deployable plugin for the HiGHmed Data Sharing Framework (DSF) [12]. It was tested using sample data across three MeDIC organizations. The important steps of this process, the sharing and aggregation of distributed cohort sizes, utilize secure Multi-Party Computation techniques in order to render a TTP superfluous.

2. State of the art

As the task of distributed feasibility queries is a common and necessary step in many medical research endeavors, various platforms that address this question exist. For example, the Clinical Communication Platform implemented by the German Cancer Consortium (DKTK) provides a central search function to request case numbers across the members' patient databases [13]. These requests await approval by the local use and access committees and are released via locally deployed software components. The results, however, disclose the number of patients on an organizational level.

The German Centre for Cardiovascular Research (DZHK) operates an architecture with an orthogonal approach by their Clinical Research Platform, providing a searchable central data repository [14].

In MII's Collaboration on Rare Diseases (CORD-MI) the MPC based analysis tool EasySMPC[3] was developed, aiming for a no-code solution. Its GUI driven usage is well suited for physician-led one-off analyses, however, its application for pipeline integration is limited.

One recent solution, able to meet most requirements, is the HiGHmed Data Sharing Framework (DSF) [12]. It uses a decentralized task queueing system and a process engine based on the open standards HL7 Fast Healthcare Interoperability Resources (HL7 FHIR R4)[4] and Business Process Model and Notation (BPMN 2.0)[5]. These components are deployed at every participating organization – the HL7 FHIR Endpoint as a publicly reachable authentication, authorization, and task queueing system and the Business Process Engine (BPE) in the internal network to execute the requested processes and communicate with local services, e.g., patient data repositories, master patient indices, or consent management services. The existing HiGHmed feasibility process [15] allows decentralized feasibility queries with optional consent checking and optional record linkage. It uses a TTP for data aggregation and record linkage purposes. A similar system, based on the HiGHmed DSF as well, is employed in the network for clinical medicine's (NUM) Covid-19 project CODEX [16].

While the DSF based solutions fulfil most of the given requirements, all of them fail to provide a high level of privacy protection for small local cohort sizes, by requiring a TTP, which might be dealing with few patient data posing a re-identification risk.

In this work, we develop, implement, and test a distributed process for the HiGHmed DSF which enables researchers to perform feasibility queries without a TTP, hence increasing both the privacy level for small local cohort sizes and the overall privacy level by utilizing mathematically provable cryptographic techniques for data protection.

3. Concept

Andrew C. Yao's seminal work [17] started the field of MPC in 1986. It was considered a theoretical technique, until the introduction of the "Fairplay" compiler [18] in 2004 and advancements in computing hardware and protocol optimizations allowed practical applications. Since then, MPC is an active research field, enabling an ever-increasing number of use cases to perform computations over distributed data sets in a privacy-

[3] https://github.com/prasser/easy-smpc
[4] https://www.hl7.org/fhir/R4
[5] https://www.omg.org/spec/BPMN/2.0

preserving manner. In principle, every calculation that is achievable using a TTP is achievable *without* a TTP using MPC protocols. However, the performance of MPC protocols is often multiple orders of magnitude slower than plain text analyses. The process is based on an extension of the GMW protocol [19] (named after its authors Goldreich, Micali, and Wigderson) to algebraic rings in order to calculate the total cohort size. Both variants work on secret shares, i.e., the secret input data is "broken up" into two or more parts. These shares do not contain any information in themself. To reconstruct the secret value, *all* shares must be recombined. Even with only one share missing no recombination is possible. By representing the computation functionality as an Arithmetic Circuit consisting of additions and multiplications, any (bounded) computation can be performed.

As feasibility queries only require the addition of values, we can exploit the additive homomorphic property of the arithmetic shares to design a comparatively simple communication protocol, suitable for the implementation into the task- and business process based DSF. The usual bottleneck in MPC performance, the available network bandwidth, does not pose a restriction for feasibility queries, as only small messages need to be transmitted.

Figure 1 illustrates the BPMN model, developed for this task. The topmost pool represents the coordinating organization, the two other pools show subprocesses executed at every participating organization. The final cohort size is calculated in an interactive protocol, the arithmetic shares are reconstructed at the coordinating organization, thus revealing the result. The complete TTP-less feasibility query process consists of the following steps:

First, a researcher defines his feasibility query at the leading organization by providing inclusion and exclusion criteria for cohort size calculation as well as whether consent checking should be performed. Currently the targets of the feasibility query process include all organizations belonging to a consortium in order to mitigate the attack scenario of disclosing individual organizations cohort sizes by performing successive feasibility queries, differing in only one excluded organization.

After the feasibility query has been created, two requests are sent to each participating organization, starting two different subprocesses. Note that both subprocesses, the lower two pools in Figure 1, are logically sequential, dividing the feasibility query execution in two stages. However, to handle network latencies and other artifacts in distributed, concurrent executions, two simultaneously executed subprocesses are involved with one waiting for the results of the other.

The logical first stage consists of the subprocess displayed in the lowest pool. After various validity checks against local and global constraints, the feasibility query is executed. If the researcher indicated to perform consent checking, the query is modified before execution to collect the Patient Identifier (PID) for each queried patient. These PIDs are used, to query the local Policy Decision Point (PDP), whether access to the patient's data has been restricted. Both cases, consent checking or not, result in the local cohort size, which is then *secret shared* (as explained above). For each participating organization one share is generated. One is held locally, the others are transmitted to all other organization, initiating the second stage, illustrated in the middle pool. Note that by withholding one share, no other participating organization can extract any information from the received shares.

In the second stage, all participating organizations wait to receive the respective shares from all other participants. If one or more organization fails to send their share, the process times out and terminates, as there is no possibility to maintain computational

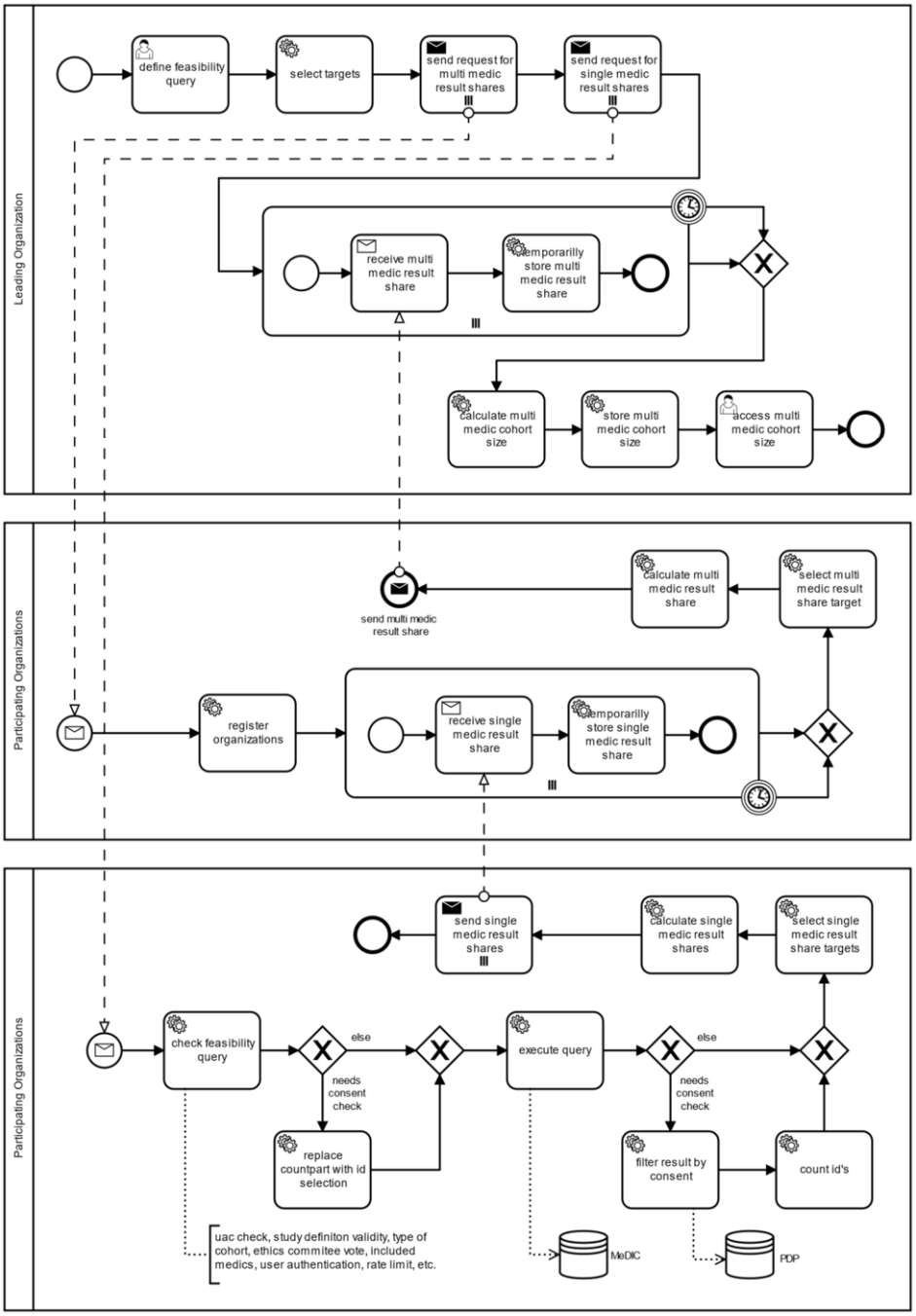

Figure 1. BPMN process diagram of the MPC feasibility process with optional consent checking

correctness with incomplete sets of shares. After all shares have been received, each organization combines the received shares, including their own, withheld one. This combination uses the homomorphic properties of the secret shares: adding all up and performing a modulo operation with the size of the used arithmetic ring, generating a valid secret share of the sum, i.e., of the desired total cohort size. This multi-MeDIC result share is sent to the leading organization, ending the computation.

In the last step, the leading organization awaits the multi-MeDIC result shares from all participating organizations. Upon receiving all of them, the shares can be recombined to reveal the clear text result, the total cohort size. The computation itself is secure against malicious adversaries, i.e., corrupted parties can only prohibit the correct calculation of the result (by injecting wrong input data or failing to send their shares), but not gain any information regarding the other parties' inputs.

4. Implementation

The proposed MPC feasibility process was implemented as a plugin process for the HiGHmed DSF in the Java programming language, using HL7 FHIR R4 resources as data model as well as BPMN 2.0 as process model, in order to remain agnostic of an organization's data repository choices and to establish semantic interoperability. The data and process model specific implementations, such as the profiled FHIR resources ResearchStudy, Group and Task, do not differ from those used in the feasibility process using a TTP for data aggregation [15]. Therefore, we would like to refer the interested reader to [15] for detailed explanations and focus on the employed cryptographic primitives and protocols in the following paragraphs.

For the secret sharing scheme, we chose a ring size of $r = 2^{32} - 1$ to fit all shares in a 32-bit integer type. As we expect cross-organizational cohort sizes of less than around 4.3 billion for virtually all use cases, this size is sufficient. However, a variant using Java's BigInteger data type is implemented, allowing arbitrary ring sizes and values. Each party in a computation with n participants generates n shares by sampling $n - 1$ uniformly independent and identically distributed Integers: $s_i \leftarrow_{\$} \{0,1,2, \dots, 2^{32} - 1\}$. The last share is generated by mixing the secret value v with all previously (randomly) generated shares, such that $v = \sum_i s_i \bmod r \; \forall i$. Due to the modular arithmetic on the algebraic ring, the last share, even though containing the secret value, is indistinguishable from a random value.

Without loss of generality, consider an addition between two parties p_1 and p_2, wanting to securely add their secret inputs v^1 and v^2, respectively. The sampled randomness during secret sharing is denoted as $rand^1$ and $rand^2$. The parties withheld one share and exchanged the other, say s_2^i. Both parties add their locally held shares, that is for p_1: $S^1 = s_1^1 + s_2^2 \; mod \; r = rand^1 + (v^2 - \mathrm{rand}^2) \; mod \; r$ and for p_2: $S^2 = s_1^2 + s_2^1 \; mod \; r = rand^2 + (v^1 - rand^1) \; mod \; r$. If they both now recombine these new two shares $R = S^1 + S^2 \; mod \; r = rand^1 + (v^2 - \mathrm{rand}^2) + rand^2 + (v^1 - rand^1) \; mod \; r = v^2 + v^1 \; mod \; r$, the randomness cancels out and the clear text sum is revealed. Of course, the two-party case is purely instructional, as knowledge of the result and the own secret value always allows to calculate the secret value of the other party.

The process was tested on three DSF instances representing three different organizations belonging to the same consortium, each containing a small data set of synthetic patient data. The open source code can be found on GitHub[6].

5. Lessons learned (Discussion)

All five previously defined requirements were met by the purposed MPC feasibility query solution. As a research advancement based on the HiGHmed DSF feasibility process, existing advantages such as fully decentralized computation without central components were retained while simultaneously advancing the state-of-the-art by significantly raising the patients' data privacy level and addressing the additional requirements of new use cases, mainly the elimination of the TTP.

The usage of open standards simplifies the integration of local systems, including the translation of cohort queries into a suitable format for each repository. This was an important consideration, as HiGHmed organizations employ a local OpenEHR repository, while NUM institutions may incorporate i2b2[7] data warehouses besides FHIR stores. Using FHIR resources as a linking data model, we provide semantic interoperability and built-in audit capabilities.

Allowing researchers to optionally perform consent checks enables additional use cases, like consent-less epidemiological studies, and creates concise interfaces for the adoption of changing legal consent requirements. In comparison to the TTP-based feasibility query process, this work does currently not support the incorporation of Record Linkage (RL). Solutions for MPC-based TTP-less RL were developed in the HiGHmed consortium [20]. The complexity and performance requirements pose a future challenge when direct integration of RL within the DSF is required. Furthermore, the development of bidirectional communication interfaces is an interesting research possibility for future work.

As no central components are employed and authorization between organizational DSF instances are handled on a pipeline- and framework level, only local user and process authorization is required to perform user authentication. This enables organizations to deploy and integrate authorization and authentication solutions of their own choosing. The complexity of inter-domain user management is avoided.

As MPC only provides input privacy, the output might pose a privacy risk. One example was already described in the "Concept" section, extracting an organization's individual cohort size by performing multiple queries, excluding one organization at a time. We mitigated this specific attack vector by forbidding the selection of individual target organizations. Other attack scenarios might be mitigated using, e.g., rate-limiting. In all cases, an audit trail is maintained by the DSF, thus adversarial behavior is identifiable.

Testing was performed on a setup with three DSF instances representing three organizations operating on small, synthetically generated data sets. While this assures correct operation, optimizing parameter values e.g., timeout durations and retry counts, must be dealt with in real, operating systems. The choice of these parameters are heavily influenced by network- and bandwidth settings, as well as the deployed hardware and firewall specifications.

[6] https://github.com/highmed/highmed-processes

[7] https://www.i2b2.org

This work is intended as a pragmatic starting point to introduce MPC protocols and analysis processes into real-world applications. It solves a real-world demand not achievable with traditional distributed computation techniques, while maintaining a reasonable scope.

6. Conclusion

To provide a decentralized feasibility process for calculating multi-centric cohort sizes with highest data privacy guarantees and without a Trusted Third Party, this work proposes a secure Multi-Party Computation based implementation using the open standards BPMN 2.0 and HL7 FHIR R4. The solution is provided as a plugin process to the HiGHmed Data Sharing Framework, allowing the easy usage for all organizations employing the HiGHmed and MII infrastructure. The process is based on the principle of data minimization by avoiding central components and, additionally, allows (optional) consent validation procedures. Identifying data never leaves organizational boundaries and data privacy regulations are acknowledged, even for small (local) cohort sizes, i.e., in rare disease research. By providing a freely available implementation under a permissive open source license, this process can be used outside the HiGHmed consortium and is easily adaptable to specific use cases.

Declarations

Conflict of Interest: The authors declare that there is no conflict of interest.

Contributions of the authors: RW and TK designed, implemented, and tested the processes. HH and CF managed the HiGHmed DSF integration and maintenance. MD and KH led and supervised the project. All authors contributed to the manuscript and substantively revised it. All authors read and approved the final manuscript.

Acknowledgements: This project is funded by the German Federal Ministry of Education and Research (BMBF, grant ids: 01ZZ1802A, 01ZZ1802E, and 01ZZ1802G). It was co-funded by the Deutsche Forschungsgemeinschaft (DFG) – SFB 1119 CROSSING/236615297. Many thanks to Fabian Prasser, Ulrich Sax, and Jürgen Eils for our fruitful discussions. The authors would like to thank all committers that contributed to the open source reference implementation and to test the current release.

References

[1] M. Grossglauser and H. Saner, “Data-driven healthcare: from patterns to actions,” *Eur J Prev Cardiolog*, vol. 21, no. 2_suppl, pp. 14–17, Nov. 2014, doi: 10.1177/2047487314552755.

[2] A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge, “Health intelligence: how artificial intelligence transforms population and personalized health,” *npj Digital Med*, vol. 1, no. 1, Art. no. 1, Oct. 2018, doi: 10.1038/s41746-018-0058-9.

[3] H. Hsin *et al.*, “Transforming Psychiatry into Data-Driven Medicine with Digital Measurement Tools,” *npj Digital Med*, vol. 1, no. 1, Art. no. 1, Aug. 2018, doi: 10.1038/s41746-018-0046-0.

[4] G. Carra, J. I. F. Salluh, F. J. da Silva Ramos, and G. Meyfroidt, "Data-driven ICU management: Using Big Data and algorithms to improve outcomes," *Journal of Critical Care*, vol. 60, pp. 300–304, Dec. 2020, doi: 10.1016/j.jcrc.2020.09.002.

[5] L. V. Rasmussen, "The Electronic Health Record for Translational Research," *J. of Cardiovasc. Trans. Res.*, vol. 7, no. 6, pp. 607–614, Aug. 2014, doi: 10.1007/s12265-014-9579-z.

[6] P. Knaup, T. M. Deserno, H.-U. Prokosch, and U. Sax, "Implementation of a National Framework to Promote Health Data Sharing," *Yearb Med Inform*, vol. 27, no. 1, pp. 302–304, Aug. 2018, doi: 10.1055/s-0038-1641210.

[7] N. Yüksekogul, N. Meyer, S. Aguduri, A. Merzweiler, and O. Heinze, "ETL-Processes for a medical data integration Center–First experiences from the heidelberg university hospital, 64," *Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS), DocAbstr*, vol. 112, 2019.

[8] S. Aguduri, A. Merzweiler, N. Yüksekogul, N. Meyer, A. Brandner, and O. Heinze, "Modeling clinical data transformation for a medical data integration center: An openEHR approach, 64," *Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS), DocAbstr*, vol. 113, 2019.

[9] T. Wendt et al., "Prozessmodelle des Data Sharing im Rahmen der Medizin- informatik-Initiative," AG Data Sharing MI-I, unpublished, 2019.

[10] Taskforce Datenschutz der MII mit Vertretern der Konsortien DIFUTURE, HiGHmed, MIRACUM, SMITH, dem Sprecher der AG Datenschutz der TMF sowie Vertretern der TMF-Geschäftsstelle, "Übergreifendes Datenschutzkonzept der Medizininformatik-Initiative," TF Datenschutz MI-I, unpublished, 2021.

[11] B. Haarbrandt *et al.*, "HiGHmed – An Open Platform Approach to Enhance Care and Research across Institutional Boundaries," *Methods Inf Med*, vol. 57, no. S 01, pp. e66–e81, May 2018, doi: 10.3414/ME18-02-0002.

[12] H. Hund, R. Wettstein, C. M. Heidt, and C. Fegeler, "Executing distributed healthcare and research Processes–The HiGHmed data sharing framework," *Studies in Health Technology and Informatics*, vol. 278, pp. 126–133, 2021.

[13] M. Lablans, E. E. Schmidt, and F. Ückert, "An Architecture for Translational Cancer Research As Exemplified by the German Cancer Consortium," *JCO Clinical Cancer Informatics*, no. 2, pp. 1–8, Dec. 2018, doi: 10.1200/CCI.17.00062.

[14] "Clinical Research Platform: DZHK." https://dzhk.de/en/research/clinical-research/clinical-research-platform/ (accessed Feb. 16, 2022).

[15] R. Wettstein, H. Hund, I. Kobylinski, C. Fegeler, and O. Heinze, "Feasibility queries in distributed Architectures–Concept and implementation in HiGHmed," in *German medical data sciences: Bringing data to life*, IOS Press, 2021, pp. 134–141.

[16] "CODEX | Netzwerk Universitätsmedizin." https://www.netzwerk-universitaetsmedizin.de/projekte/codex (accessed Feb. 16, 2022).

[17] A. C. Yao, "How to generate and exchange secrets," in *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*, Oct. 1986, pp. 162–167. doi: 10.1109/sfcs.1986.25.

[18] D. Malkhi, N. Nisan, B. Pinkas, and Y. Sella, "Fairplay — A Secure Two-Party Computation System," 2004, p. 17.

[19] O. Goldreich, S. Micali, and A. Wigderson, "How to Play ANY Mental Game," New York, NY, USA, 1987. doi: 10.1145/28395.28420.

[20] S. Stammler *et al.*, "Mainzelliste SecureEpiLinker (MainSEL): Privacy-Preserving Record Linkage using Secure Multi-Party Computation," *Bioinformatics*, 2020, doi: 10.1093/bioinformatics/btaa764.