DISSERTATION

Innovative methods for sharing data across institutions in medical research

Innovative Verfahren für die standortübergreifende Datennutzung in der medizinischen Forschung

zur Erlangung des akademischen Grades
Doctor rerum medicinalium (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Felix Nikolaus Wirth

Erstbetreuer: Univ.-Prof. Dr. Fabian Prasser

Datum der Promotion: 30. Juni 2024

Table of contents

List of tables

List of figures

List of abbreviations

CORD_MI	Collaboration on rare diseases
GDPR	General Data Protection Regulation
GMW	Goldreich, Micali and Wigderson
GUI	Graphical User Interface
HIPAA	Health Insurance Portability and Accountability Act
IMAP	Internet Message Access Protocol
ms	Milliseconds
MB	Megabyte
SMPC	Secure Multiparty Computation
SMTP	Simple Mail Transfer Protocol
US	United States

Zusammenfassung

Moderne datengetriebene medizinische Forschungsansätze („Künstliche Intelligenz“, „Data Science“) benötigen große Datenmengen („Big Data“). Dies kann im Regelfall nur durch eine institutionsübergreifende Datennutzung erreicht werden („Data Sharing“). Datenschutz und der Schutz der Privatsphäre der Betroffenen ist dabei eine zentrale Herausforderung. Um dieser zu begegnen, können verschiedene Methoden, wie etwa Anonymisierungsverfahren oder föderierte Auswertungen, eingesetzt werden. Allerdings findet Data Sharing in der Praxis nur selten statt, obwohl es von vielen Seiten gefordert und gefördert wird. Ein Grund hierfür ist die Unklarheit über Vor- und Nachteile verschiedener Data Sharing-Ansätze. Erstes Ziel dieser Arbeit war es, ein Instrument zu entwickeln, welches diese Vor- und Nachteile transparent macht. Das Instrument bewertet Ansätze anhand von zwei Dimensionen - *Nutzen* und *Schutz* - wobei jede Dimension mit drei Achsen weiter differenziert ist. Die Achsen bestehen etwa aus dem Grad des Schutzes der Privatsphäre, der durch die Ergebnisse der durchgeführten Analysen gewährleistet wird oder der Flexibilität einer Plattform hinsichtlich der Arten von Analysen, die durchgeführt werden können. Das Instrument wurde zu Evaluationszwecken für die Analyse des Status Quo sowie zur Identifikation von Lücken und Potenzialen für innovative Verfahren eingesetzt. Als zweites Ziel wurde anschließend ein innovatives Werkzeug für den praktischen Einsatz von kryptographischen Data Sharing-Verfahren entwickelt. Der Einsatz entsprechender Ansätze scheitert bisher vor allem an zwei Barrieren: (1) der technischen Komplexität beim Aufbau einer Kryptographie-basierten Data Sharing-Infrastruktur und (2) der Benutzerfreundlichkeit kryptographischer Data Sharing-Verfahren, insbesondere für medizinische Forschende. Das neue Werkzeug EasySMPC zeichnet sich dadurch aus, dass es eine kryptographisch sichere Berechnung von Summen (beispielsweise Häufigkeiten von Diagnosen) über Institutionsgrenzen hinweg auf Basis einer einfach zu bedienenden graphischen Benutzeroberfläche ermöglicht. Zur Anwendung ist weder technische Expertise noch der Aufbau spezieller Infrastrukturkomponenten notwendig. Die Praxistauglichkeit von EasySMPC wurde in einer ausführlichen Performance-Evaluation experimentell analysiert.

Abstract

Implementing modern data-driven medical research approaches ("Artificial intelligence", "Data Science") requires access to large amounts of data ("Big Data"). Typically, this can only be achieved through cross-institutional data use and exchange ("Data Sharing"). In this process, the protection of the privacy of patients and probands affected is a central challenge. Various methods can be used to meet this challenge, such as anonymization or federation. However, data sharing is currently put into practice only to a limited extent, although it is demanded and promoted from many sides. One reason for this is the lack of clarity about the advantages and disadvantages of different data sharing approaches. The first goal of this thesis was to develop an instrument that makes these advantages and disadvantages more transparent. The instrument systematizes approaches based on two dimensions - utility and protection - where each dimension is further differentiated with three axes describing different aspects of the dimensions, such as the degree of privacy protection provided by the results of performed analyses or the flexibility of a platform regarding the types of analyses that can be performed. The instrument was used for evaluation purposes to analyze the status quo and to identify gaps and potentials for innovative approaches. Next, and as a second goal, an innovative tool for the practical use of cryptographic data sharing methods has been designed and implemented. So far, such approaches are only rarely used in practice due to two main obstacles: (1) the technical complexity of setting up a cryptography-based data sharing infrastructure and (2) a lack of user-friendliness of cryptographic data sharing methods, especially for medical researchers. The tool EasySMPC, which was developed as part of this work, is characterized by the fact that it allows cryptographically secure computation of sums (e.g., frequencies of diagnoses) across institutional boundaries based on an easy-to-use graphical user interface. Neither technical expertise nor the deployment of specific infrastructure components is necessary for its practical use. The practicability of EasySMPC was analyzed experimentally in a detailed performance evaluation.

1. Introduction

In order to use recent medical research approaches ("Artificial intelligence ", "Data Science") access to large amounts of data is necessary ("Big Data") [1]. Typically, this can only be achieved through cross-institutional data use and exchange ("Data Sharing") [2]. Data sharing is being promoted by various organizations [3–5], has been described as an “ethical and scientific imperative” [6] and it is expected to be a standard practice in the future. However, it is important to distinguish between different types of data sharing: (1) data sharing in the sense of publishing raw data, for example together with scientific articles for reproducibility purposes, and (2) data sharing to pool data across institutions to improve sample sizes [7]. This thesis puts a specific focus on the latter challenge and the term data sharing will be used accordingly.

Although the scientific community as well as the public have a positive attitude towards data sharing, it has not yet become widely established in practice [8]. A prominent obstacle to practical data sharing are laws restricting how medical data can be processed e.g., for privacy protection reasons [9]. The US Health Insurance Portability and Accountability Act (HIPAA) [10] and the EU General Data Protection Regulation (GDPR) [11] are two important examples. Moreover, the willingness of patients and citizens to share their own data has been shown to be much higher when privacy is maintained [12].

Different methods have emerged for providing a solid legal basis for performing data-driven research, also in cross-institutional settings: (1) obtaining *informed consent* by affected patients and participants, (2) *anonymization of data*, which refers to the process of changing data so that it cannot be traced back to specific individuals, or (3) the use of more complex *data sharing infrastructures*, which have specifically been designed to support the privacy-preserving analysis of data which is distributed across multiple institutions (for a more detailed description of (2) and (3) see below). A typical example in the context of data sharing infrastructures would be the exchange of aggregated, non-personal data (see Section 3.1 for further details on different types of data sharing infrastructures). While obtaining informed consent can be considered the gold standard from the data protection and ethics perspective, it is often not possible if data is to be used in retrospect on a large scale. Moreover, anonymization is challenging for high-dimensional data [13]. As a consequence, a range of alternative approaches has been developed, which exhibit different strengths and weaknesses. The central importance of privacy-preserving data sharing technologies and the need to develop them

further is clearly underlined by the recently published National Strategy to Advance Privacy-Preserving Data Sharing and Analytics of the US government [14]. I believe that the resulting heterogeneity of the proposed approaches to data sharing is one of the reasons why they are not widely used in practice.

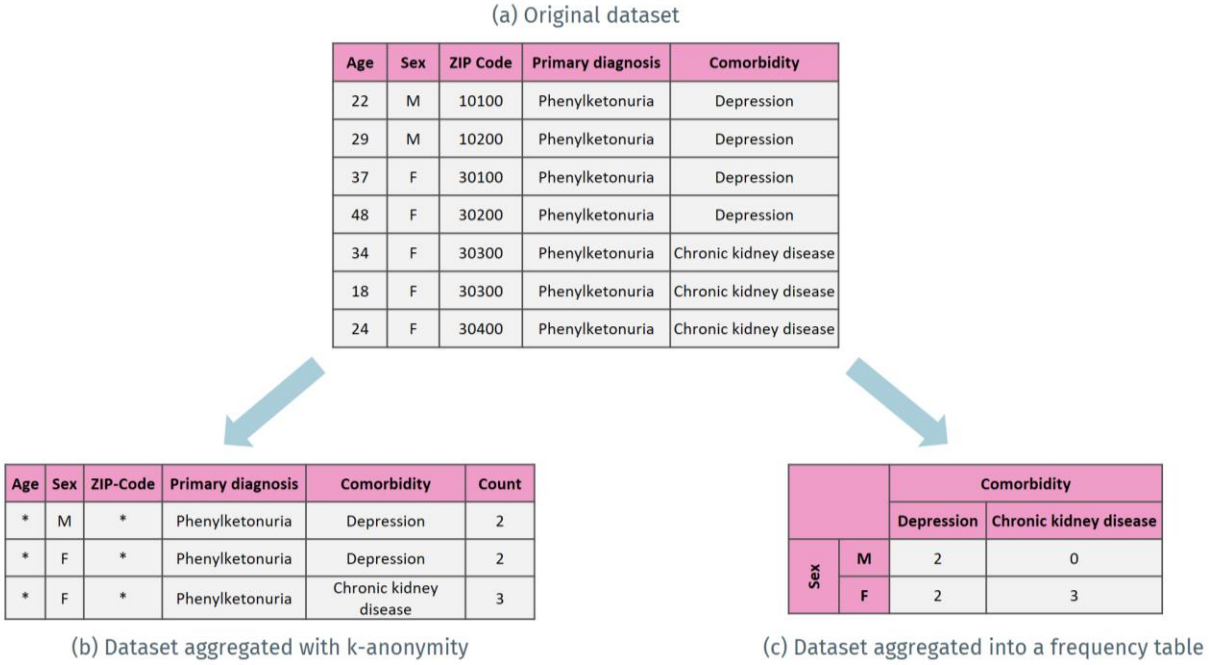
1.1 Risks associated with sharing different types of medical data

As mentioned in the previous section, the main motivation for developing more complex data sharing infrastructures is a legal need to ensure the anonymity of the data subjects that often arises. Intuitively, one might believe that this can be achieved by removing directly identifying data, such as names and addresses from data. However, this approach was disproved by the famous case of William Weld, in which structured medical data about the then-governor of Massachusetts was re-identified in an alleged anonymous dataset by combining it with publicly available information [15]. A wide variety of research has shown that this is possible with different types of medical data, such as genetic data [16], clinical free text [17] and medical images [18]. The focus of this work is, however, structured tabular data. There are several ways in which the anonymity and privacy of individuals can be breached based on data. They can be systematized into different categories [19]:

- *Membership disclosure:* The possibility of an attacker to learn whether data referring to a certain individual is in a dataset or not [20].
- *Attribute disclosure:* The possibility of an attacker to learn about sensitive attributes of an individual [21]. Please note that this does not necessarily mean that the specific record of the individual is identified.
- *Identity disclosure:* The possibility of an attacker to link one or several records in a dataset to an individual [22].

As the example of William Weld already shows, protecting data from such threats while sharing it with others is challenging. One simple and popular approach for protecting individual-level data is performing data anonymization using k-anonymity [23], in which data is aggregated into groups of not less than k indistinguishable individuals. However, as already noted above, this process quickly reaches its limits when data is high-dimensional or contains information about a small number of individuals [24]. At the same time, scientific analyses usually aim for statistical results and hence aggregated data.

While this can provide some degree of protection, additional privacy controls are also needed for aggregated data, as is illustrated in Figure 1.



(a) Original dataset

Age	Sex	ZIP Code	Primary diagnosis	Comorbidity
22	M	10100	Phenylketonuria	Depression
29	M	10200	Phenylketonuria	Depression
37	F	30100	Phenylketonuria	Depression
48	F	30200	Phenylketonuria	Depression
34	F	30300	Phenylketonuria	Chronic kidney disease
18	F	30300	Phenylketonuria	Chronic kidney disease
24	F	30400	Phenylketonuria	Chronic kidney disease

Age	Sex	ZIP-Code	Primary diagnosis	Comorbidity	Count
*	M	*	Phenylketonuria	Depression	2
*	F	*	Phenylketonuria	Depression	2
*	F	*	Phenylketonuria	Chronic kidney disease	3

(b) Dataset aggregated with k-anonymity

		Comorbidity	
		Depression	Chronic kidney disease
Sex	M	2	0
	F	2	3

(c) Dataset aggregated into a frequency table

Figure 1: An example of a protection of a dataset with k-anonymity and with aggregation (own illustration).

In subfigure (a), the figure shows an example of an original dataset for the rare disease Phenylketonuria with some demographic data about patients as well as a documented comorbidity. I point out that the dataset is a simplified example intended for illustrative purposes only. In subfigure (b), the figure shows an example of a modified version of the dataset fulfilling k-anonymity with k = 2. The aggregation provides a certain level of privacy, but also reduces the utility of the data as information has been removed. However the dataset still contains a potential privacy thread: an attacker knowing that data of a specific individual male can be found in the dataset can easily infer that this named individual must suffer from depression as well, which constitutes attribute disclosure.

Another example for the problem is the possibility of retrieving personal information from frequency tables when cell counts are low or zero [25]. This problem is illustrated in part (c) of Figure 1 with an example that is closely related to the example in part (b). The subfigure displays a frequency table which has been generated for the variables sex and comorbidity. The table shows the number of common occurrences of the corresponding values of the variables. Since this is aggregated data, one might think that it poses no privacy risk to the data subjects. However, as with sub-figure (b), an attacker who knows

about the presence of a particular men in the dataset can infer that this men suffers from depression, since all men in the dataset have this comorbidity.

Similar problems can also arise with other forms of aggregated data [26]. The issue has been formalized in the *Database Reconstruction Theorem* stating that with too many, too accurate statistics generated about a protected dataset the original dataset can be reconstructed [27]. Other well-known examples for the general problem include (1) possible membership disclosure attacks using the p-values of a genetic statistical analysis as long as the genome is known [28], (2) the possible reconstruction of the original data when only co-variance values are known [29] or (3) possible membership attacks from machine learning model parameters [30]. However, it is worthwhile noting that the data privacy related risks induced by the described problems can be reduced when increasing the amount of data processed. Part (c) of Figure 1 can serve as an example for this claim: An increase of the number of patients displayed could lead to a frequency table in which no cell has a count of zero. This would in turn not allow for the above described attribute disclosure.

1.2 Aims and contributions of this work

The previous sections highlight the challenges of preserving privacy while sharing data and the wide range of methods that have been proposed for this purpose, leading to significant complexity in assessing the landscape of available solutions. Consequently, this thesis approaches the topic in two consecutive steps. First, I hypothesized that a systematic categorization of properties of privacy-preserving data sharing methods can successfully map diverse implementations into a unified framework, providing a coherent perspective on the current landscape of solutions. Second, I hypothesized that a "no-code" cryptographic data sharing tool that requires no dedicated technical setup is feasible, thus addressing a significant barrier to adoption in medical environments identified in my first contribution.

The instrument to systematize and compare data sharing approaches was designed to make their advantages and disadvantages more transparent. To this end, the instrument categorizes approaches based on two dimensions - utility and protection - where each dimension is further differentiated along three axes. The instrument was then used to analyze the status quo and to identify gaps and potentials for innovative approaches.

The innovative “no-code” tool EasySMPC for cryptographic data sharing in medicine has been designed to (1) work without requiring a specific technical setup in hospitals and (2) be usable for medical researchers without programming knowledge (cf. the recent recommendation of the US government to “improve usability and inclusiveness of PPDSA [privacy-preserving data sharing and analytics] solutions” [14] as a strategic priority to promote data sharing). The feasibility of EasySMPC was analyzed experimentally in a detailed performance evaluation.

The development and evaluation of these two methods constitute the contribution of this dissertation. Their details will be presented in the chapters below.

2. Methods

In this section, I will briefly present the methods developed during my research. The structure of the section reflects my approach of working on the topic from two sides that also build upon on each other. The first method has been developed for improving the ability to evaluate and compare medical data sharing approaches. The second method aims to overcome certain limitations of existing approaches and can thus enable new ways of sharing data.

2.1 Development of a novel method to assess data sharing infrastructures

2.1.1 Theoretical framework

Research in medical informatics, which is a field on the intersection of computer science and medicine, can be performed with a variety of methods. Vessey, Ramesh, and Glass proposed a taxonomy with 19 different classes of research methods in computer science, including (1) *proof of concept implementation*, (2) *action research, (3) conceptual analysis* and *(4) simulation* [31]. Since the goal of the first part of this work was to systematize and analyze data sharing infrastructures, I performed a conceptual analysis. This is closely related to *conceptualization* which is defined as creating an “abstract, simplified view of the world that we wish to represent for some purpose” [32]. The conceptual analysis was performed following the research framework proposed by Holz et al. [33]. Here, the idea is to structure the research process along four key questions: (A) “What do we want to achieve?”, (B) “Where does the data come from?”, (C) “What do we do with the data?” and (D) “Have we achieved our goal?”.

Figure 2: Process implemented while researching for the method (adopted from [33]).

Figure 2 illustrates the application of the framework while developing a systematization for assessing data sharing infrastructures. As can be seen, my aim was to design a systematization (A), which was derived from data about existing data sharing approaches (B) that was then clustered (C). Finally, the systematization was validated by using it to compare different approaches and identify gaps (D).

2.1.2 Existing concepts and methods

Ideally, data sharing could simply be realized by loading all relevant data into a central, potentially cross-institutional database, with which data scientists could interact to perform research. This approach would provide a very high degree of scientific usefulness. However, it offers limited privacy protection and is often not possible, if anonymity guarantees are needed to obtain a legal basis for processing (see Section 1.1). Any measures implemented to further improve privacy protection will inevitably lead to a reduced scientific utility. This can be visualized analogously to the well-known “risk-utility curves”, which are often used to study data anonymization methods [34], as is illustrated in Figure 3.

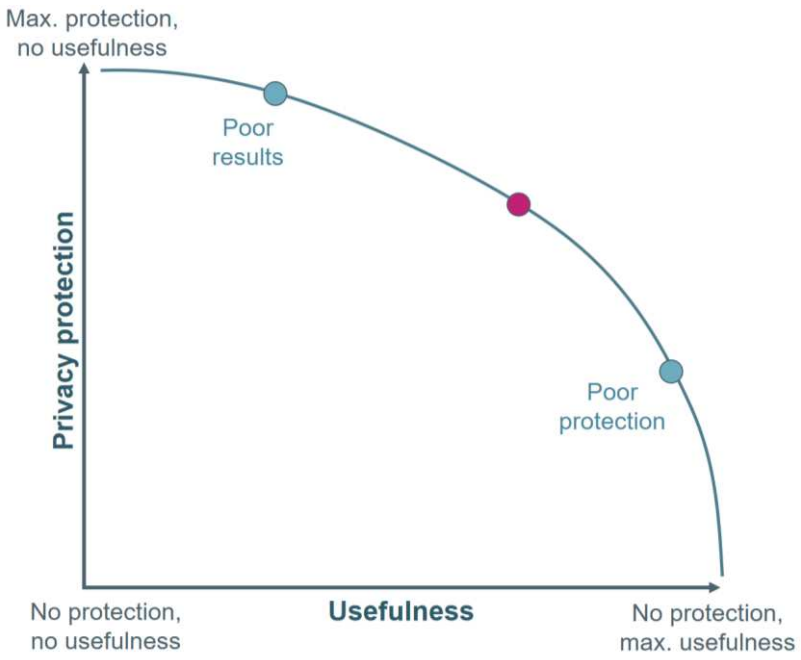


Figure 3: Relationship of privacy protection and data usefulness (adopted from [35])

As can be seen, privacy protection needs to be balanced against scientific usefulness. The point “poor results”, for example, indicates a data sharing solution protecting privacy very well but failing to provide accurate analytical results (e.g., based on anonymization), while the point “poor protection”, indicates an approach providing accurate analytical results but little privacy protection (e.g., the central database mentioned above).

For reasoning about the protection provided by offering secure access to a central database, the well-known Five Safes Framework has been proposed [36], which is illustrated in Figure 4.

Figure 4: Axes of the Five Safes Framework [35].

As the name indicates, the framework covers five different aspects that define the degree of protection provided during data access:

- *Safe People*: Captures the degree of trustworthiness of researchers who are provided with data access.
- *Safe Project*: Reflects the appropriateness (e.g., from a legal and ethical perspective) of the projects which are carried out with the data.

- *Safe Data*: Covers the degree of identifiability of the data that is accessed.
- *Safe Settings*: Captures the degree of protection provided through the access mechanism including access rules and roles.
- *Safe Output:* Reflects the privacy risks associated with statistical results generated during data use.

2.1.3 A novel systematization to assess data sharing infrastructures

This work proposes a novel method in form of a framework and systematization for studying the trade-off between protection and usefulness provided by data sharing approaches, analogously to risk-utility curves for anonymization methods. In the process of creating the systematization, I selected the three technical aspects of the Five Safes Framework – which has been designed to reason about protective measures used to safeguard access to a central database – and described how they can be used to describe properties of a wider variety of data sharing infrastructures. Moreover, I defined three utility aspects representing common and important requirements for biomedical research projects [37]. Finally, I combined all aspects into a holistic framework. An overview of the systematization is shown in Figure 5.

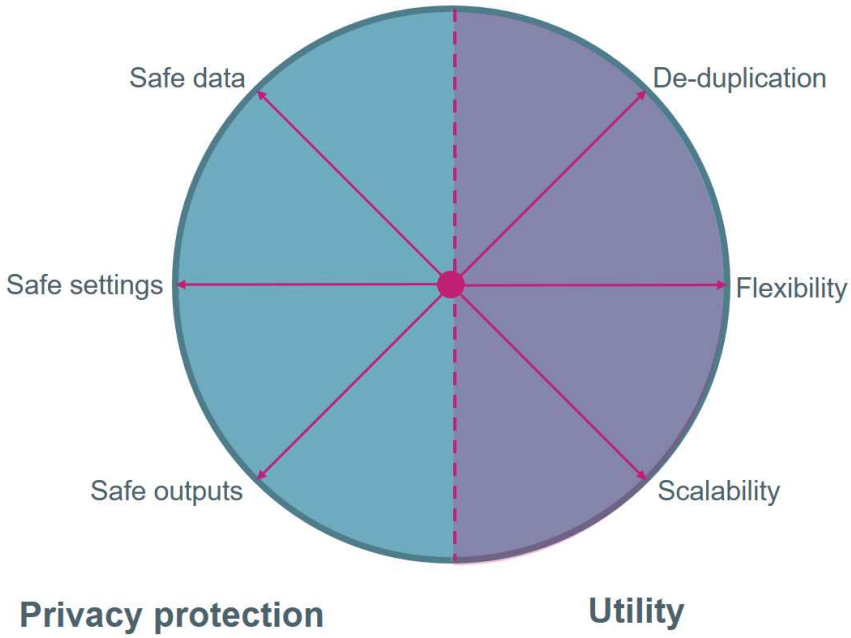


Figure 5: Dimensions and axes of the developed framework (adopted from [35]).

The figure illustrates the privacy protection dimension on the left and the usefulness dimension on the right with three axes, i.e. aspects, each. As mentioned, the axes for the degree of privacy protection are derived from the Five Safes Framework:

- *Safe Data*: The approach addresses privacy risks on the data level, e.g., by applying anonymization, aggregation or encryption.
- *Safe Settings*: The approach addresses privacy risks through secure environments, in which the users have limited access to the data.
- *Safe Outputs*: The approach ensures that results produced do not result in privacy risks.

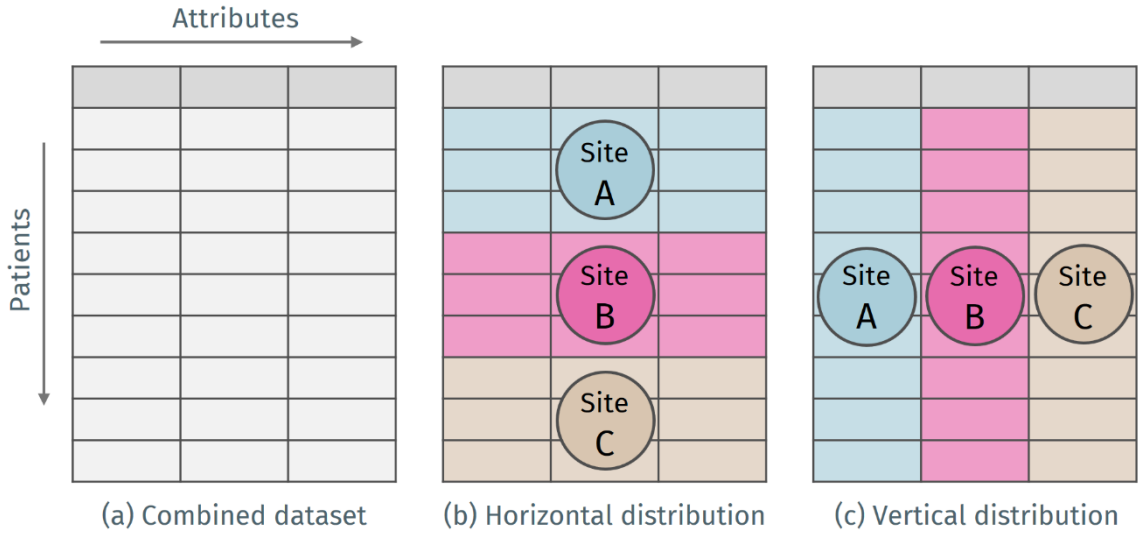


Figure 6: Horizontal and vertical data distribution [35].

Moreover, I derived three axes describing the usefulness of data sharing approaches from common requirements in real-world data-driven research:

- *De-duplication / Record Linkage:* This axis describes the ability of an approach to support (1) linkage between vertically distributed data or (2) identification of an overlap in patient populations for vertically distributed data. As illustrated in
  Figure 6, data can be distributed *horizontally* (i.e. records about patients from the population are distributed across different sites, cf.
  Figure 6b) or *vertically* (i.e. different properties about the same patients or probands can be distributed across different sites, cf.
  Figure 6c). Combinations of horizontal and vertical distribution are also common in practice.

- *Flexibility*: This axis describes the degree to which an approach can be used to implement different types of analyses or whether it is extendable with new analytical methods.
- *Scalability*: This axis covers the scalability of an approach, e.g., in terms of execution times, storage requirements or hardware requirements as well as required network bandwidth and latency when increasing sites, data volume or dimensionality.

2.2 Design and development of a novel data sharing method

2.2.1 Theoretical framework

The second part of the work described in this thesis focused on filling one gap regarding cryptographic-based data sharing infrastructures that was identified by applying the previously described method (cf. Section 4.1.3). Referring to the taxonomy by Vessey, Ramesh, and Glass, I used the following methods: (1) *software product implementation* and (2) *laboratory experiments with software* for evaluation purposes. Those steps were combined in a cycle following the framework by Holz et al. [33], as is illustrated in Figure 7.

Figure 7: Research framework for designing and developing a novel data sharing method (adopted from [33]).

As can be seen, a software product (A) was designed and implemented (B), then evaluated (C) and tuned in an iterative process until acceptable performance was achieved (D).

A particular focus was put on developing a usable product, as several authors stress the importance of developing software products and not only prototypes in research, since this provides more reliable evidence on the practicability of a new method [38–40]. Given the fact that one prominent goal of medical informatics is to support biomedical research with new findings *and* good software systems, development is a crucial part of research in medical informatics [41].

2.2.2 Secure multiparty computation

The general idea behind the cryptographic methods of *Secure Multiparty Computation* can be described with the following analogy [42]: Confidential data of different parties is shared with a trusted third party. This third party performs computations (e.g., statistical analyses) and shares the results with the parties involved. Therefore, no input data of any party is shared with another party (apart from the trusted third party). SMPC protocols provide the same guarantees without the need of a trusted third party by means of cryptographic methods. A very common example is the “millionaires' problem” in which two millionaires want to find out who is richer without revealing their bank account statements to one another. Yao proposed *garbled circuits* to solve this problem [43]. Another famous protocol is the *GMW protocol* developed by Goldreich, Micali and Wigderson [44], in which a Boolean function can be computed with the secure inputs of several parties by using logical XOR and AND operations. GMW can be extended to an *Arithmetic Secret Sharing* protocol, in which a common sum of values from three or more parties can be calculated, without sharing the single summands.

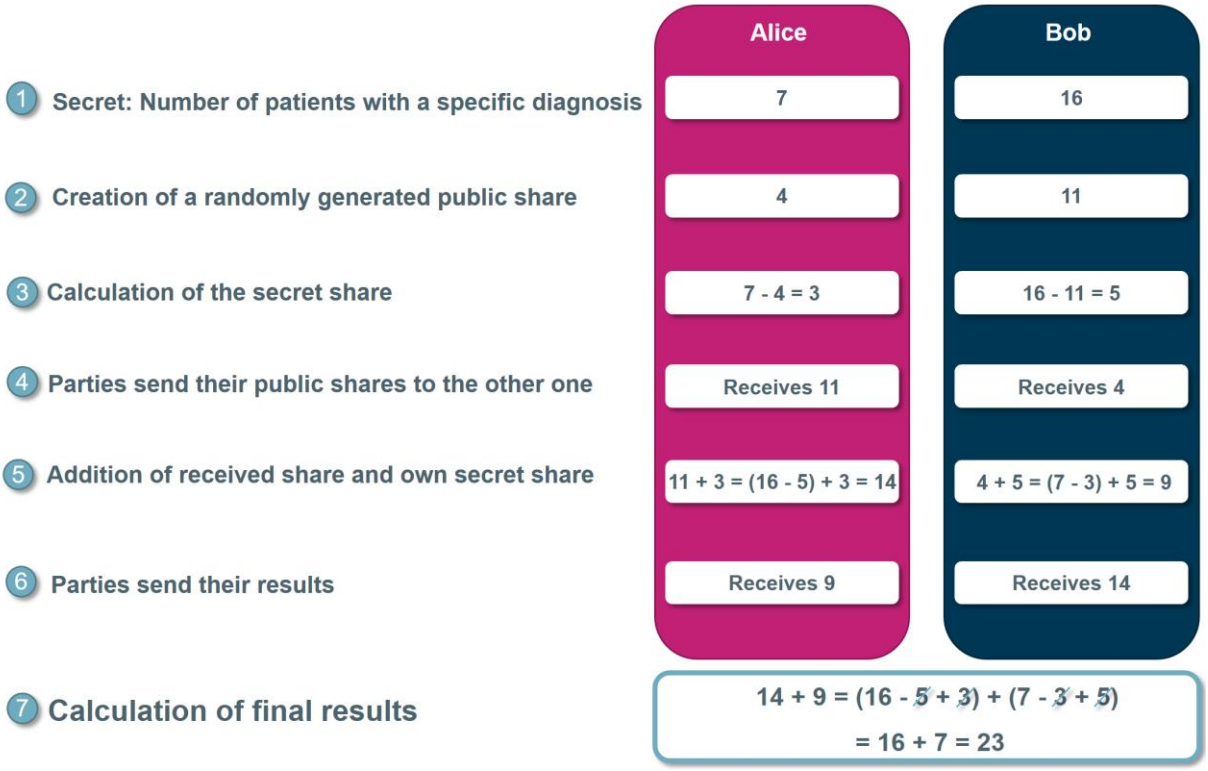


Figure 8: Example of the Arithmetic Secret Sharing protocol executed with two parties (own illustration).

Figure 8 shows an exemplary, simplified execution of the Arithmetic Secret Sharing protocol with two participants. Both participants “Alice” and “Bob” have a secret number they want to sum up without revealing their own number to one another. To achieve this, they each generate a random number, which is the public share. The secret share is the difference between the secret number and the public share. The secret share will remain confidential and the public share is sent to the respective other party in a first round of communication. The received public share will then be added with the secret share and this sum will be sent again to the respective other party in a second round of communication. The last line shows that, at the end of the protocol, the shares cancel each other out so that only the sum of the two secret original values remains. I note that an actual use of the protocol always requires at least three participants to keep the secret values confidential and that real-world implementations of the protocol are a bit more complex. Moreover, I note that this approach is suitable for exchanging aggregated data (that in fact often is privacy-sensitive as well; see Section 1.1) and not for sharing individual-level data. For a detailed discussion of analyses supported and limitations of the approach I refer to Section 4.2.3.

2.2.3 The software EasySMPC

The Arithmetic Secret Sharing protocol is a good first step for performing basic statistical analysis in a secure manner across sites for biomedical research. This is especially true for research on rare diseases in which population numbers are so low that other types of secure data sharing mechanisms usually fail to protect privacy.

However, as the analysis presented in Section 3.1.2 showed, SMPC-based methods like Arithmetic Secret Sharing have not been widely adopted in practice. One important reason is that cryptographic methods are not very approachable to non-IT-experts. To improve this, I designed and implemented EasySMPC, an innovative software focusing on usable SMPC. The software supports the steps illustrated in Figure 8 through a user-friendly graphical interface, supporting the secure addition and subtraction of data from three or more participants (e.g., hospitals). It has been developed as a cross-platform, stand-alone application and was designed for non-technical users. EasySMPC is implemented in the programming language Java following the widely-used model-view-controller pattern [45]. The model-view-controller pattern separates the responsibility of the different software modules into (1) a data model, (2) the view presented to the user and (3) a controller interacting with the model and the view. EasySMPC runs on the local computers of the participants (e.g., physicians in a hospital) and installation is supported through an easy-to-use installer. No additional infrastructure, such as server backends are needed, since EasySMPC uses e-mails as its communicative backbone. The most important modules of the software are shown in Figure 9.

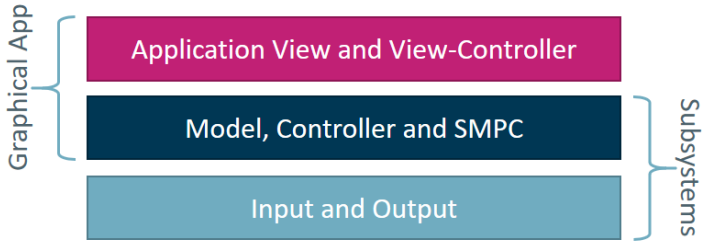


Figure 9: Architecture of EasySMPC [46].

As the figure illustrates, the Application View and View-Controller module is built on top of two subsystems for (1) SMPC operations as well as data manipulation and (2) interaction with external applications as well the other participants. In more detail, the three modules serve the following purposes:

- *Application View and View-Controller*: The module consists of eight different perspectives developed with Swing, a programming library for developing Graphical User Interfaces (GUIs) in Java, directing the user through the different steps of the process.
- *Model, Controller and SMPC*: This module includes the application's data model allowing controlled access to and manipulation of data during the execution of the protocol. Moreover, the module includes the implementation of the Arithmetic Secret Sharing protocol.
- *Input and Output*: The module provides functionalities to import and export data in different formats (e.g., Excel, text files). Moreover, the module is responsible for sending and receiving e-mail messages automatically.

The implementation relies on Java standard libraries as well as Jakarta Mail and the packages "POI", "Commons" and "Logging" provided by the Apache Software Foundation.

2.2.4 Performance evaluation

Since EasySMPC is, to the best of my knowledge, the first software to use e-mail as a communication channel for running SMPC protocols and SMPC protocols are well known to have significant requirements in terms of network performance, I conducted an extensive evaluation to study the behavior of the software. To cover a wide range of application scenarios, I varied two technical and two user-specific factors as part of the evaluation:

- *Technical factor 1: Polling frequency* - The frequency at which EasySMPC checks for new e-mail messages (settings used: 1, 5, 10, 15 and 20 seconds).
- *Technical factor 2: Network latency* - The time data packages in a network take from the sender to the receiver network. This was simulated using the tool *tc*[1] (settings used: 30 milliseconds (ms) to simulate data exchange in a national project and 100 ms to simulate data exchange in an overseas setup).
- *User-specific factor 1: Number of participants* - The number of participants (e.g., hospitals) taking part in the computation (settings used: 3, 5, 10 and 20).

[1] https://man7.org/linux/man-pages/man8/tc.8.html

- *User-specific factor 2: Number of variables* - The number of variables summed up within one calculation (settings used: 1000, 2500, 5000 and 10000).

The performance evaluation was conducted on a single computer with 64 1.8 GHz AMD EPYC 7502 CPUs and 512 GB of RAM running CentOS 8.4 as an operation system and a dedicated Mailserver (iRedMail) installed. The evaluation was executed 15 times for each of combination of user-specific and technical factors described above. The following outcome variables were collected: (1) the time needed to complete an EasySMPC calculation process, (2) the number of messages sent and (3) the overall data volume exchanged. The code and the experimental results are publicly available [47].

3. Results

In this section, I will briefly present the results of my research. In line with my research approach, the results consist of a conceptual and a methodological contribution.

3.1 Assessment and comparison of data sharing infrastructures

3.1.1 Principal results

The developed systematization was applied to assess various data sharing infrastructures, such as DataSHIELD [48], OMOP/OHDSI [49] or the Personal Health Train approach [50]. The results are illustrated in Table 1. The table shows the respective infrastructure, the year in which a first paper describing the approach was published, its classification along the different axes as well as the category assigned through clustering (see Section 3.1.2).

Table 1: Results of the analysis of solutions for privacy-preserving data sharing [35]

Approach	Year	Category	1. Privacy protection			2. Usefulness		
			1. Safe Data	2. Safe Setting	3. Safe Outputs	1. De-Duplication	2. Flexibility	3. Scal-ability
SHRINE/i2b2	2008	Distributed data analysis	Yes	No	Yes[b]	No	Low	Yes
DataSHIELD	2010	Distributed data analysis	Yes	No	Yes[b]	No	High	Yes
OHDSI	2014	Distributed data analysis	Yes	No	Yes[b]	No	High	Yes
Personal Health Train	2017	Distributed data analysis	Yes	No	Yes[b]	No	High	Yes
Clinerion	2015	Distributed data analysis	Yes	No	Yes[b]	No	Low	Yes
TriNetX	2015	Distributed data analysis	Yes	No	Yes[b]	No	Low	Yes
MedCo	2018	Secure multiparty computation	Yes[a]	Yes	Yes	No	Low	No
Sharemind	2008	Secure multiparty computation	Yes[a]	Yes	Yes	Yes	High	No
Scottish national Safe Haven	2015	Data enclave	No	Yes	Yes	Yes	High	Yes

Virtual Research Center	2014	Data enclave	No	Yes	Yes	Yes	Low	Yes

[a] The processed data is encrypted individual-level data and thus safe.

[b] *Safe Outputs* is an implicit result of providing *Safe Data* as input.

3.1.2 Categories of data sharing infrastructures identified

Based on shared properties of the different solutions listed in Table 1, I were able to group the results into three different categories of data sharing infrastructures:

- *Distributed data analysis*: This category refers to solutions exchanging only aggregated (and hence non-personal) data. A common example is the calculation of a mean value locally for each participating party, which is thereafter exchanged with the other parties. The assessed solutions provided *Safe Data* and *Safe Outputs*, while *Safe Settings* are not necessary, since *Safe Data* is provided already as input. Regarding scientific usefulness, none of the assessed infrastructures from this category allowed for *De-Duplication / Record-Linkage* and all provided a high degree of *Scalability*. *Flexibility* is usually high, although there are limitations as to what can be calculated by exchanging aggregated data only. A typical example for a *Distributed Data Analysis* approach is implemented in the software DataSHIELD [48].
- *Secure Multiparty Computation*: This category refers to solutions using SMPC protocols. These approaches only exchange encrypted *Safe Data*. The cryptographic algorithms forming the basis can also be understood as providing a *Safe Setting*. It must be noted SMPC protocols don't necessarily provide *Safe Outputs*, as they can reveal information about individuals if not designed and used carefully. For instance, a SMPC protocol might output a statistical table with small cell counts, which can make individuals identifiable (see Section 1.1). Moreover, *Flexibility* as well as *Scalability* are a challenge for SMPC-based solutions, while support for *De-Duplication / Record-Linkage* can be provided by some implementations. A typical example for an SMPC-based infrastructure is the software MedCo [51].
- *Data enclaves*: This category refers to solutions in which data is submitted to a trusted-third party, who allows eligible researchers to perform analyses against the data through safe access methods, e.g., monitored remote desktop connections. Therefore, no *Safe Data* is processed and stored in the enclave.

However, the data is processed in a *Safe Setting* and usually measures are implemented to ensure that only *Safe Output* can be exported. Moreover, *Data Enclaves* support *De-Duplication / Record-Linkage* and provide *Scalability,* while the *Flexibility* of the assessed solutions differed. A typical example for a *Data Enclave* is the *US Center for Medicare and Medicaid Services Virtual Research Data Center* [52].

3.2 A new tool enabling cryptography-based data sharing

3.2.1 Software overview

EasySMPC is an easy-to-use software for users with little or no technical knowledge. The user is guided through the secure data sharing process based on six consecutive perspectives, of which four are shown in Figure 10.

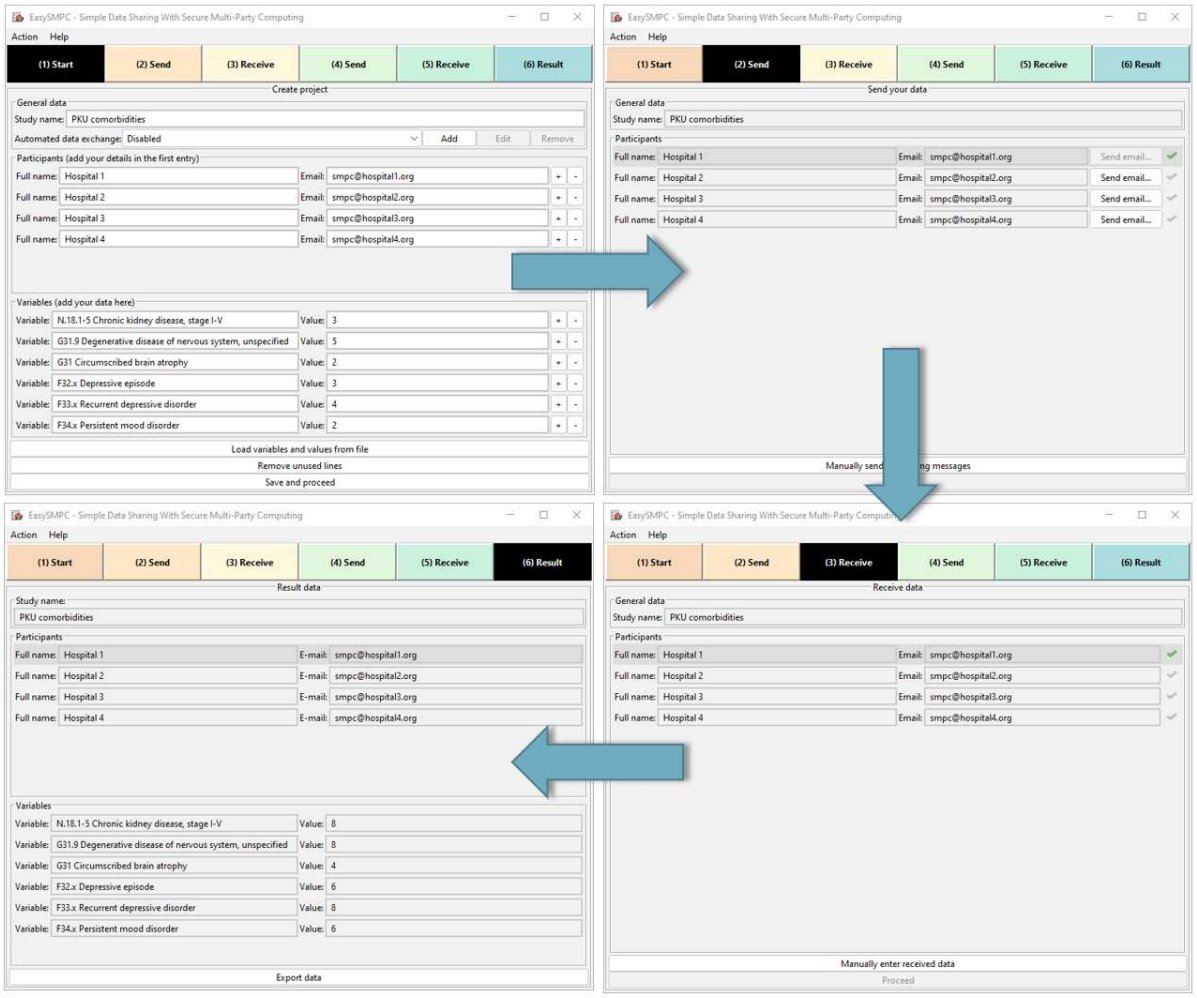


Figure 10: Perspectives in EasySMPC [46].

In the first perspective the study can be set up and contact details for the individual sites can be added. Moreover, the data can either be entered manually or loaded from Excel or CSV files. Please note that the participants' contact details as well as the names of the variables will be shared with the other participants while the data values entered next to the names remain confidential. In the second perspective the data is send to all sites for the first round of communication (cf. Section 2.2.2 for a description of the rounds). In the third perspective data is received for the first round of communication. In the fourth and fifth perspective, the processes executed in the second and third perspective are repeated to perform the second round of communication. They are thus omitted in Figure 10. Finally, in the last perspective the results are displayed.

Sites are identified by email addresses and all data is exchanged via the Simple Mail Transfer Protocol (SMTP) and the Internet Message Access Protocol (IMAP). The software also features a dialog for connecting it to mailboxes.

3.2.2 Performance evaluation

In this section, I present results of the performance evaluation with a network latency of 30 ms (national data sharing scenario). For further results I refer the interested reader to the respective publication [46]. Figure 11 provides an overview of the data volumes exchanged with different settings.

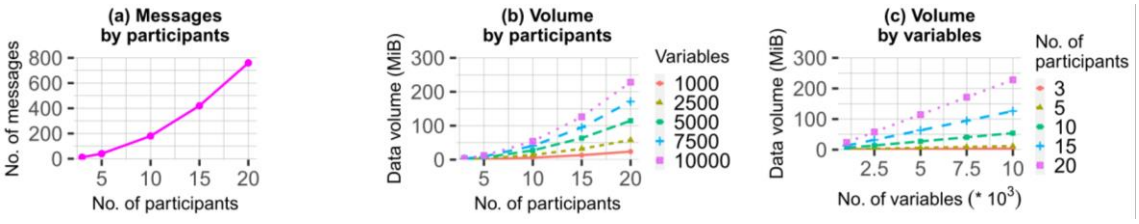


Figure 11: Number of messages and total data volume exchanged (30 ms) [46].

Figure 11a shows that the number of messages exchanged by EasySMPC grows quadratically with an increasing number of sites involved in a computation. The total data volume increases analogously with the number of participating sites (Figure 11b). The figure further shows a linear influence of the number of variables on the exchanged data volume (Figure 11c). Overall execution times are illustrated in Figure 12 for different polling intervals, number of variables and number of participating sites.

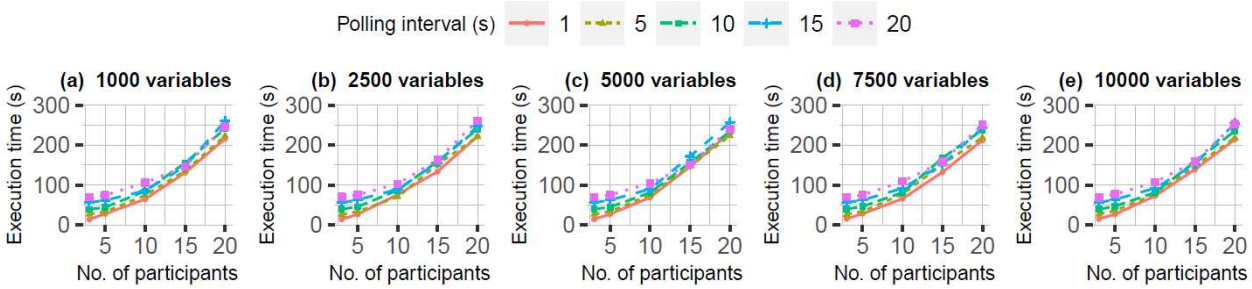


Figure 12: Execution times for increasing numbers of participants and variables as well as different polling frequencies (30 ms) [46].

As can be seen, execution times correlate with the data volume exchanged and grows quadratically with the number of participating sites and linearly with the number of variables. Increasing the polling frequency reduces execution times.

In summary, the numbers clearly show that EasySMPC is scalable enough to be used in real-world applications. Summing up the value of 10,000 variables over 20 sites from the same geographic area (e.g., country) can be performed in under 5 minutes.

4. Discussion

In this section, I will briefly discuss and compare the results of my work. The structure of the section reflects the two directions from which I have worked on the topic of medical data sharing. The next section (see "Conclusions"), will bring both contributions together.

4.1 Discussion of the method to assess data sharing infrastructures

4.1.1 Principal results

A method in form of a systematization was developed to capture and assess the characteristics of data sharing approaches. The systematization consists of two dimensions each containing three different axes. Moreover, the systematization was used to assess ten existing data sharing approaches showing its practicability. By clustering approaches based on their properties, three different categories of approaches with similar characteristics could be identified. It is notable that the category *Distributed Data Analysis* was most prominent. One explanation for this might be the fact that *Distributed Data Analysis* infrastructures are relatively simple to develop, explain and use. *SMPC*-based infrastructures as well as *Data Enclaves* are either more complex in terms of technology or are associated with additional legal challenges. Based on the systematization and the analysis of existing approaches, I was able to identify important gaps in the system landscape, such as a lack of user-friendly SMPC-based solutions.

4.1.2 Comparison with prior work

The presented systematization method builds upon the Five Safes Framework [36], which has been used for assessing data access environments in various areas of research. Moreover, other work has been published assessing data sharing activities in the biomedical area, which, however, differ from my approach: For instance, Knoppers [53] proposed a framework focusing on trust, compliance and responsible research, while Aziz et al. [54] gave an overview of data sharing with a specific focus on genomic data and cryptographic methods. Similarly, Mittos et al. [55] present a framework to asses technologies providing privacy protection when working with genomic data in various scenarios, not only when sharing data. Moreover, Thapa et al. [56] published an overview of data sharing specific to the area of „precision health“, also focusing on different cryptographic techniques. To the best of our knowledge, the proposed systematization is

the first targeting the trade-off between privacy and usefulness for different types of data sharing approaches for biomedical research with different types of data.

4.1.3 Limitations and future work

A limitation of the presented systematization is its qualitative nature and its suitability only for high-level comparisons. Although a more formalized and detailed framework would be desirable, it is very challenging to create such a framework. Some reasons for this are given by Richie and Green when arguing for the qualitative nature of the Five Safes Framework [57]. They argue, for instance, that it is very difficult to model the interactions between the different *Safes* correctly. Moreover, a formalized and quantitative framework would require precise metrics for data privacy and usefulness, which is still an open research problem (a paper by Wagner and Eckoff [58] listed over 80 different privacy models suggested by the research community). Thus, there are several potential directions for future work: Firstly, the presented systematization could be extended, for example by incorporating the most common privacy models and by adding more axes, such as *User-Friendliness*. Moreover, there is the potential to bridge open gaps, such as the lack of user-friendly SMPC-based approaches. Current SMPC-based solutions often require complex infrastructures and service-side components to set up, leading to hurdles for getting them deployed in real-world settings and operated securely. In addition to that, SMPC-based solutions often rely on software libraries or domain-specific languages and require knowledge of and experience with programming for their application. I have already worked towards a solution that does not come with such requirements in the second part of my doctoral work.

4.2 Discussion the new data sharing approach

4.2.1 Principal results

I have designed and developed EasySMPC, which supports an innovative method for the secure calculation of sums across different sites while keeping the single summands confidential. The tool is user-friendly and easy to roll-out, as no server infrastructure or network setup is necessary. EasySMPC uses existing e-mail infrastructures. The results of our performance evaluation demonstrate its applicability in real-world contexts. EasySMPC is available as open source software and the source code

as well as comprehensive documentation, executables and installers ca be obtained online [59].

4.2.2 Comparison with prior work

Several other SMPC-based data sharing solutions have been proposed. The solutions can be roughly assigned to four different categories:

- *Record-linkage processes:* Some related works describe the secure conduction of record-linkage processes with SMPC protocols [60,61].
- *Specific usage cases*: Other related works describe SMPC-based solutions for specific use cases in biomedical research, such as drug discovery [62], genome-wide association studies [63] or genomic diagnostics [64].
- *Specific statistical method*s: Further related works describe the secure implementation of specific statistical methods, such as Kaplan-Meier estimators [65] or regression analyses [66].
- *Generic frameworks and data sharing infrastructures*: Finally, some related works present SMPC-based generic programming frameworks like Sharemind MPC [67], ABY [68] or FRESCO [69] and complete SMPC-based data sharing platforms such as MedCo [51] or FAMHE [70].

In comparison to the first three groups of related work, EasySMPC is rather generic and focused on users with a non-technical background. Moreover, as described before, it is an actual software product and not only a research prototype. This also applies to other solutions such as MedCo and FAMHE. However, the deployment and operation of these products requires much more effort and technical knowledge.

4.2.3 Limitations and future work

EasySMPC is a tool for exchanging aggregated data securely across sites to generate common aggregated statistics. Typical applications include the calculation of descriptive statistics of cohorts with certain characteristics across sites, such as the overall prevalence of a condition or sex or age distributions (see Section 4.2 in [46] for an overview of statistical methods supported). I emphasize that specific methods are needed for calculating such statistics across sites if the characteristics are rare in a cohort (cf. Section 1.1). For example, under many circumstances, counts of less than 11 are considered to be personal information under the GDPR [71] and hence restrictions apply

to how this data can be shared with others. The goal of EasySMPC is to enable performing such analyses with as little hurdles as possible. On the legal side this is achieved by exchanging and combining data in encrypted form only. While there is still some legal uncertainty around the question of whether this constitutes a processing of personal data, the simplicity of the algorithms employed by EasySMPC facilitate legal assessments. On the technical side EasySMPC strives to require as few preparations as possible, while on the user side a point-and-click user interface makes it as easy as possible to participate in joint calculations. Ultimately, I hope that this will help to get SMPC-based data sharing in medicine into broader use. However, to calculate some statistics with EasySMPC, manual steps are necessary.

EasySMPC also has two limitations regarding privacy protection and security: Firstly, since the participants only authenticate with their e-mail addresses, a man-in-the-middle-attack could be performed, in which an intruder replaces a regular participant. However, this attack could only lead to wrong results at the end of the calculation and not to the disclosure of data from another site. Secondly, when considering the privacy protection dimension of the systematization developed in the first part of my doctoral project, EasySMPC provides *Safe Data* and a *Safe Setting*, but not necessarily *Safe Outputs.* As with other SMPC-based approaches, care needs to be taken to ensure that anonymity is provided. Future work could extend EasySMPC with more robust authentication methods and integrate process-level anonymization processes, such as Differential Privacy [72], to also protect its outputs.

5. Conclusions

Bringing both methods and their results together, Figure 13 shows an assessment of EasySMPC with the new systematization.

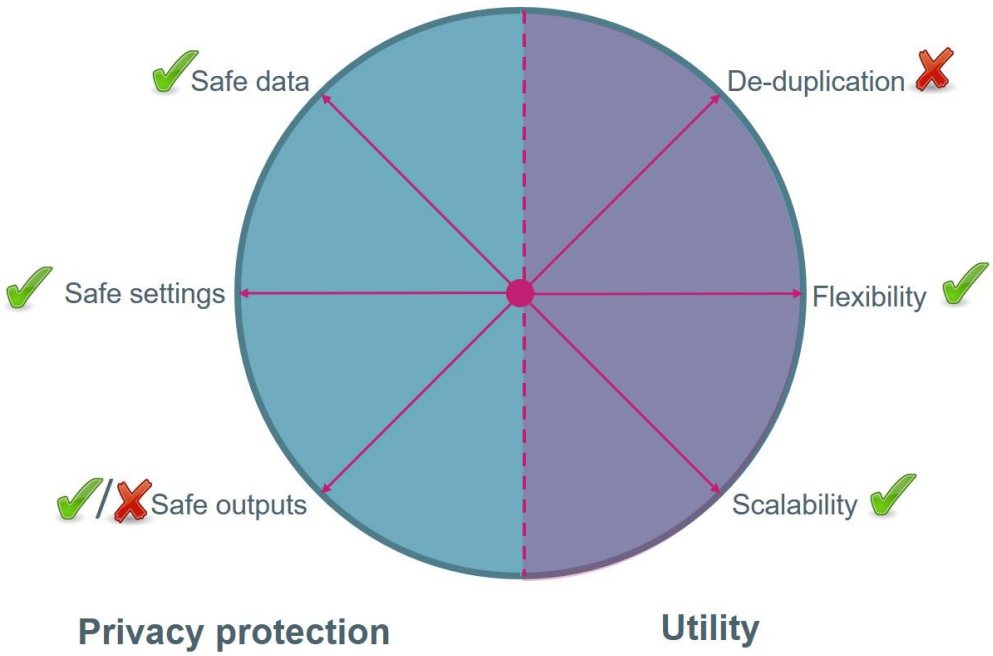


Figure 13: An assessment of EasySMPC with the developed systematization (adopted from [35]).

As can be seen, EasySMPC offers two *Safes* in the privacy protection dimension as well as *Scalability* and *Flexibility* in the utility dimension. Whether EasySMPC provides *Safe Outputs* depends on the input utilized. If, for example, every participating site but one provides a value of zero for a variable, the non-zero value of one site will be disclosed. This can be prevented in different ways, however. One option is to utilize a minimum threshold rule to ensure that each site contributes a non-zero value or using an additional round of EasySMPC to determine whether enough sites are able to contribute non-zero values. The user-friendliness of EasySMPC is currently not reflected by the systematization.

Recently, a push can be observed towards the real-world deployment of privacy-preserving data sharing infrastructures in Germany. For example, DataSHIELD has been used in the Medical Informatics Initiative [73] and, based on specific laws created for this purpose, *Data Enclaves* are being established in the Centre for Cancer Registry Data and in the Research Data Center at the Federal Institute for Drugs and Medical Devices. Also EasySMPC is being used in practice in the project “Collaboration on Rare Diseases”

(CORD_MI). Since the disease investigated in CORD_MI are rare by definition, it is hard to collect and aggregate epidemiological data on them without running into privacy challenges. EasySMPC is being used in CORD_MI to overcome these challenges while being able to focus on legal and scientific - not technical - questions. While EasySMPC is only suitable for sharing aggregated statistical data, this already enables a range of questions to be answered. For example, the CORD_MI project is interested in understanding how many women suffering from cystic fibrosis gave birth to a child between the years 2015 and 2022 across all German university hospitals. This can be answered with EasySMPC while ensuring that the individual site-specific statistics are not disclosed.

Reference list

[1] Munevar S. Unlocking big data for better health. Nat Biotechnol. 2017;35:684–6. Available from: https://doi.org/10.1038/nbt.3918.

[2] Hulsen T. Sharing Is caring - data sharing initiatives in healthcare. Int J Environ Res Public Health. 2020;17. Available from: https://doi.org/10.3390/ijerph17093046.

[3] Pilat D, Fukasaku Y. OECD principles and guidelines for access to research data from public funding. Data Sci J. 2007;6:OD4–11. Available from: https://doi.org/10.2481/dsj.6.OD4.

[4] Carr D, Littler K. Sharing Research Data to Improve Public Health. J Empir Res Hum Res Ethics. 2015;10:314–6. Available from: https://doi.org/10.1177/1556264615593485.

[5] Australian Government - National Health and Medical Research Council. Open Access Policy 2018. Available from: https://www.nhmrc.gov.au/file/15242/download?token=rgNjnh0B.

[6] Bauchner H, Golub RM, Fontanarosa PB. Data sharing: An ethical and scientific imperative. JAMA. 2016;315:1238–40. Available from: https://doi.org/10.1001/jama.2016.2420.

[7] Institute of Medicine (US). Sharing Clinical Research Data: Workshop Summary. Washington: The National Academies Press; 2013.

[8] Gabelica M, Bojčić R, Puljak L. Many researchers were not compliant with their published data sharing statement: a mixed-methods study. J Clin Epidemiol. 2022;150:33–41. Available from: https://doi.org/10.1016/j.jclinepi.2022.05.019.

[9] Vis DJ, Lewin J, Liao RG, Mao M, Andre F, Ward RL, Calvo F, Teh BT, Camargo AA, Knoppers BM, Sawyers CL. Towards a global cancer knowledge network: dissecting the current international cancer genomic sequencing landscape. Ann. Oncol. 2017;28:1145–51. Available from: https://doi.org/10.1093/annonc/mdx037.

[10] Act A. Health insurance portability and accountability act of 1996. Public Law 1996;104:191.

[11] Regulation GDP. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. Official Journal of the European Union (OJ) 2016;59:294.

[12] Kim KK, Joseph JG, Ohno-Machado L. Comparison of consumers' views on electronic data sharing for healthcare and research. J Am Med Inform Assoc. 2015;22:821–30. Available from: https://doi.org/10.1093/jamia/ocv014.

[13] Prasser F, Eicher J, Spengler H, Bild R, Kuhn KA. Flexible data anonymization using ARX—Current status and challenges ahead. Softw - Pract Exp. 2020;50:1277–304. Available from: https://doi.org/10.1002/spe.2812.

[14] National Science and Technology Council. National strategy to advance privacy-preserving data sharing and analytics. Washington: Executive Office of the President of the United States; 2023. Available from: https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf.

[15] Barth-Jones D. The 're-identification'of governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. 2012. Available from: http://dx.doi.org/10.2139/ssrn.2076397.

[16] Heeney C, Hawkins N, de Vries J, Boddington P, Kaye J. Assessing the privacy risks of data sharing in genomics. Public Health Genom. 2011;14:17–25. Available from: https://doi.org/10.1159/000294150.
[17] Uzuner Ö, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc. 2007;14:550–63. Available from: https://doi.org/10.1197/jamia.M2444.
[18] Aryanto KYE, Oudkerk M, van Ooijen PMA. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. Eur Radiol. 2015;25:3685–95. Available from: https://doi.org/10.1007/s00330-015-3794-0.
[19] Li T, Li N, Zhang J, Molloy I. Slicing: A new approach for privacy preserving data publishing. IEEE Trans Knowl Data Eng. 2010;24:561–74. Availble from: https://doi.org/10.1109/TKDE.2010.236
[20] Nergiz ME, Atzori M, Clifton C. Hiding the presence of individuals from shared databases. In: Zhou L, Ling T, editors. SIGMOD/PODS07: Proceedings of the 2007 ACM SIGMOD international conference on management of data; 2007 June 11 – 14; Beijing. New York: Association for Computing Machinery; 2007, p. 665–76. Available from: https://doi.org/10.1145/1247480.1247554.
[21] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. ACM Trans Knowl Discov Data. 2007;1:3-es. Available from: 10.1145/1217299.1217302.
[22] Sweeney L. Computational disclosure control: A primer on data privacy protection. Dissertation. Massachusetts Institute of Technology; 2001. Available from: https://core.ac.uk/download/pdf/4393391.pdf.
[23] Sweeney L. k-anonymity: a model for protecting privacy. Int J Unc Fuzz Knowl Based Syst. 2002;10:557–70. Available from: https://doi.org/10.1142/S0218488502001648.
[24] Aggarwal CC. On k-anonymity and the curse of dimensionality. In: Bratbergsengen K, editor. VLDB '05: Proceedings of the 31st international conference on very large data bases; 2005 August 30 – September 02; Trondheim. New York: Association for Computing Machinery; 2005, p. 901–9. Available from: https://doi.org/10.5555/1083592.1083696
[25] Smith D, Elliot M. A measure of disclosure risk for tables of counts. Trans Data Priv. 2008;1:34–52. Available from: 10.5555/1556401.1556405.
[26] Denning DE, Denning PJ. The tracker: A threat to statistical database security. ACM TODS. 1979;4:76–96. Available from: https://doi.org/10.1145/320064.320069.
[27] Dinur I, Nissim K. Revealing information while preserving privacy. In: Nevem F, editor. PODS'03: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems; 2003 June 09 – 11; San Diego. New York: Association for Computing Machinery; 2003, p. 202–10. Available from: https://doi.org/10.1145/773153.773173.
[28] Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008;4. Available from: https://doi.org/10.1371/journal.pgen.1000167.
[29] Huth M, Arruda J, Gusinow R, Contento L, Tacconelli E, Hasenauer J. Accessibility of covariance information creates vulnerability in Federated Learning frameworks. BioRxiv. 2022:2022–10. Available from: https://doi.org/10.1101/2022.10.09.511497.

[30] Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: Butler KRB, editor. SP 2017: Proceedings of the IEEE Symposium on Security and Privacy; 2017 May 22 – 26; San Jose. Washington: IEEE Computer Society; 2017, p. 3–18. Available from: https://doi.org/10.1109/SP.2017.41.
[31] Glass RL, Ramesh V, Vessey I. An analysis of research in computing disciplines. Commun ACM. 2004;47:89–94. Available from: https://doi.org/10.1145/990680.990686.
[32] Gruber TR. Toward principles for the design of ontologies used for knowledge sharing? Int J Hum Comput. 1995;43:907–28. Available from: https://doi.org/10.1006/ijhc.1995.1081.
[33] Holz HJ, Applin A, Haberman B, Joyce D, Purchase H, Reed C. Research Methods in Computing: What are they, and how should we teach them? ACM SIGCSE. 2006;38(4):96–114. Available from: https://doi.org/10.1145/1189136.1189180.
[34] Li T, Li N. On the tradeoff between privacy and utility in data publishing. In: Elder J, Fogelman-Soulié F, editors. KDD: Proceedings of the 15th ACM SIG-KDD International Conference on Knowledge Discovery and Data Mining; 2009 June 28 – July 1; Paris. New York: Association for Computing Machinery; 2009, p. 517–26. Available from: https://doi.org/10.1145/1557019.1557079.
[35] Wirth FN, Meurers T, Johns M, Prasser F. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. BMC Med Inform Decis Mak. 2021;21:242. Available from: https://doi.org/10.1186/s12911-021-01602-x.
[36] Desai T, Ritchie F, Welpton R. Five Safes: designing data access for research. 2016. Available from: https://doi.org/10.13140/RG.2.1.3661.1604.
[37] Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics. 2016;17:73. Available from: https://doi.org/10.1186/s12910-016-0153-x.
[38] Winkler D, Mordinyi R, Biffl S. Research prototypes versus products: lessons learned from software development processes in research projects. In: McCaffery F, O'Connor RV, Messnarz R, editors. EuroSPI 2013: Proceedings of the 20th European Conference on Systems, Software and Services Process Improvement; 2013 June 25 - 27; Dundalk. Berlin: Springer; 2013, p. 48–59.
[39] Jacobs J. From prototype to product: deployment strategies in computer science research. XRDS. 2016;23:5–6. Available from: https://doi.org/10.1145/2983451.
[40] Odom W, Wakkary R, Lim Y, Desjardins A, Hengeveld B, Banks R. From research prototype to research product. In: Kaye J, Druin A, editors. CHI'16: Proceedings of the 2016 CHI conference on human factors in computing systems; 2016 May 7 – 12; San Jose. New York: Association for Computing Machinery; 2016, p. 2549–61. Available from: https://doi.org/10.1145/2858036.2858447.
[41] Hauschild A-C, Martin R, Holst SC, Wienbeck J, Heider D. Guideline for software life cycle in health informatics. IScience. 2022;25:105534. Available from: https://doi.org/10.1016/j.isci.2022.105534.
[42] Canetti R. Security and Composition of Multiparty Cryptographic Protocols. J Cryptology. 2000;13:143–202. Available from: https://doi.org/10.1007/s001459910006.

[43] Yao AC-C. How to generate and exchange secrets. SFCS ’86: Proceedings of the 27th Annual symposium on foundations of computer science; 1986 October 27 -29. Washington: IEEE Computer Society; 1986, p. 162–7. Available from: https://doi.org/10.1109/SFCS.1986.

[44] Micali S, Goldreich O, Wigderson A. How to play any mental game. In: Aho A, editor. STOC ’87: Proceedings of the Nineteenth ACM Symposium on Theory of Computing; 25-27 May 1987; New York. New York: Association for Computing Machinery; 1987, p. 218–29. Available from: https://doi.org/10.1145/28395.28420.

[45] Krasner GE, Pope ST. A description of the model-view-controller user interface paradigm in the smalltalk-80 system. J Op Prog. 1988;1:26–49.

[46] Wirth FN, Kussel T, Müller A, Hamacher K, Prasser F. EasySMPC: a simple but powerful no-code tool for practical secure multiparty computation. BMC Bioinform. 2022;23:531. Available from: https://doi.org/10.1186/s12859-022-05044-8.

[47] Wirth F. EasySMPC performance evaluation. 2021. Available from: https://github.com/easy-smpc/easy-smpc-performance-evaluation.

[48] Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, Minion J, Boyd AW, Newby CJ, Nuotio ML, Wilson R. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014;43:1929–44. Available from: https://doi.org/10.1093/ije/dyu188.

[49] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, Van Der Lei J. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574–8. Available from: https://doi.org/10.3233/978-1-61499-564-7-574.

[50] Beyan O, Choudhury A, van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, Karim MR, Dumontier M, Decker S, da Silva Santos LO, Dekker A. Distributed analytics on sensitive medical data: The Personal Health Train. Data Intelligence. 2020;2:96–107. https://doi.org/10.1162/dint_a_00032.

[51] Raisaro JL, Troncoso-Pastoriza JR, Misbach M, Sousa JS, Pradervand S, Missiaglia E, Michielin O, Ford B, Hubaux JP. MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data. IEEE/ACM Trans Comput Biol Bioinform. 2018;16:1328–41. Available from: https://doi.org/10.1109/TCBB.2018.2854776.

[52] ResDAC. CMS Virtual Research Data Center (VRDC). Minnesota: ResDAC; 2023. Available from: https://www.resdac.org/cms-virtual-research-data-center-vrdc-faqs

[53] Knoppers BM. Framework for responsible sharing of genomic and health-related data. Hugo J. 2014;8;3. Available from: https://doi.org/10.1186/s11568-014-0003-1.

[54] Aziz MM, Sadat MN, Alhadidi D, Wang S, Jiang X, Brown CL, Mohammed N. Privacy-preserving techniques of genomic data - a survey. Brief Bioinform. 2019;20:887–95. Available from: https://doi.org/10.1093/bib/bbx139.

[55] Mittos A, Malin B, Cristofaro ED. Systematizing Genome Privacy Research: A Privacy-Enhancing Technologies Perspective. In: Chatzikokolakis K, Troncoso C, editors. PoPETs: Proceedings on privacy enhancing technologies; 2019; July 16 - 20; Stockholm. Berlin: De Gruyter; 2019:87–107. Available from: https://doi.org/10.2478/popets-2019-0006.

[56] Thapa C, Camtepe S. Precision health data: Requirements, challenges and existing techniques for data security and privacy. Comput Biol Med. 2021;129:104130. Available from: https://doi.org/10.1016/j.compbiomed.2020.104130.
[57] Ritchie F, Green E. Frameworks, principles and accreditation in modern data management. Bristol: UWE; 2020. Available from: https://uwe-repository.worktribe.com/output/6790882/frameworks-principles-and-accreditation-in-modern-data-management.
[58] Wagner I, Eckhoff D. Technical Privacy Metrics: A Systematic Survey. ACM Comput Surv. 2018;51:57:1-57:38. Available from: https://doi.org/10.1145/3168389.
[59] Wirth FN, Kussel T, Müller A, Hamacher K, Prasser F. EasySMPC source code. 2023. Available from: https://github.com/prasser/easy-smpc.
[60] Stammler S, Kussel T, Schoppmann P, Stampe F, Tremper G, Katzenbeisser S, Hamacher K, Lablans M. Mainzelliste SecureEpiLinker (MainSEL): Privacy-Preserving Record Linkage using Secure Multi-Party Computation. Bioinform. 2022 15;38(6):1657-68. Available from: https://doi.org/10.1093/bioinformatics/btaa764.
[61] El Emam K, Samet S, Hu J, Peyton L, Earle C, Jayaraman GC, Wong T, Kantarcioglu M, Dankar F, Essex A. A Protocol for the secure linking of registries for HPV surveillance. PLoS One. 2012;7:e39915. Available from: https://doi.org/10.1371/journal.pone.0039915.
[62] Ma R, Li Y, Li C, Wan F, Hu H, Xu W, Zeng J. Secure multiparty computation for privacy-preserving drug discovery. Bioinform. 2020;36:2872–80. Available from: https://doi.org/10.1093/bioinformatics/btaa038.
[63] Bonte C, Makri E, Ardeshirdavani A, Simm J, Moreau Y, Vercauteren F. Towards practical privacy-preserving genome-wide association study. BMC Bioinform. 2018;19:537. Available from: https://doi.org/10.1186/s12859-018-2541-3.
[64] Jagadeesh KA, Wu DJ, Birgmeier JA, Boneh D, Bejerano G. Deriving genomic diagnoses without revealing patient genomes. Science. 2017;357:692–5. Available from: https://doi.org/10.1126/science.aam9710.
[65] von Maltitz M, Ballhausen H, Kaul D, Fleischmann DF, Niyazi M, Belka C, Carle G. A privacy-preserving log-rank test for the Kaplan-Meier estimator with secure multiparty computation: Algorithm development and validation. JMIR Med Inform. 2021;9:e22158. https://doi.org/10.2196/22158.
[66] Shi H, Jiang C, Dai W, Jiang X, Tang Y, Ohno-Machado L, Carle G. Secure Multi-pArty Computation Grid LOgistic REgression (SMAC-GLORE). BMC Med Inform Decis Mak. 2016;16:89. Available from: https://doi.org/10.1186/s12911-016-0316-1.
[67] Archer DW, Bogdanov D, Lindell Y, Kamm L, Nielsen K, Pagter JI, Smart NP, Wright RN. From keys to databases - real-world applications of secure multi-party computation. Comput J. 2018;61:1749–71. Available from: https://doi.org/10.1093/comjnl/bxy090.
[68] Demmler D, Schneider T, Zohner M. ABY- a framework for efficient mixed-protocol secure two-party computation. NDSS ’15: Network and distributed system security symposium; 2015 February 08 - 11; San Diego. Reston: The Internet Society; 2015, p. 497-512. Available from: https://doi.org/10.14722/ndss.2015.23113.
[69] Alexandra Institute. FRESCO - A framework for efficient secure computation 2021. Available from: https://github.com/aicis/fresco.

[70] Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, Cuendet MA, Sousa JS, Cho H, Berger B, Fellay J, Hubaux JP. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Nat Commun. 2021;12:5910. Available from: https://doi.org/10.1038/s41467-021-25972-y.

[71] European Medicines Agency. External guidance on the implementation of the European medicines agency policy on the publication of clinical data for medicinal products for human use. 2018. Available from: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-3.pdf.

[72] Dwork C. Differential Privacy: A Survey of Results. In: Agrawal M, Du D, Duan Z, Li A, editors. TAMC 2008: Theory and Applications of Models of Computation 5th International Conference; 2008 April 25 - 29; Xi'an. Heidelberg: Springer; 2008, p. 1–19. Available from: https://doi.org/10.1007/978-3-540-79228-4_1.

[73] Semler S, Wissing F, Heyder R. German Medical Informatics Initiative: A national approach to integrating health data from patient care and medical research. Methods Inf Med. 2018;57:e50–6. Available from: https://doi.org/10.3414/ME18-03-0003.

Statutory declaration

“I, Felix Nikolaus Wirth, by personally signing this document in lieu of an oath, hereby affirm that I prepared the submitted dissertation on the topic “Innovative methods for sharing data across institutions in medical research” (“Innovative Verfahren für die standortübergreifende Datennutzung in der medizinischen Forschung”), independently and without the support of third parties, and that I used no other sources and aids than those stated.
All parts which are based on the publications or presentations of other authors, either in letter or in spirit, are specified as such in accordance with the citing guidelines. The sections on methodology (in particular regarding practical work, laboratory regulations, statistical processing) and results (in particular regarding figures, charts and tables) are exclusively my responsibility.

Furthermore, I declare that I have correctly marked all of the data, the analyses, and the conclusions generated from data obtained in collaboration with other persons, and that I have correctly marked my own contribution and the contributions of other persons (cf. declaration of contribution). I have correctly marked all texts or parts of texts that were generated in collaboration with other persons.

My contributions to any publications to this dissertation correspond to those stated in the below joint declaration made together with the supervisor. All publications created within the scope of the dissertation comply with the guidelines of the ICMJE (International Committee of Medical Journal Editors; http://www.icmje.org) on authorship. In addition, I declare that I shall comply with the regulations of Charité – Universitätsmedizin Berlin on ensuring good scientific practice.

I declare that I have not yet submitted this dissertation in identical or similar form to another Faculty.

The significance of this statutory declaration and the consequences of a false statutory declaration under criminal law (Sections 156, 161 of the German Criminal Code) are known to me.”

13.10.2023

Date Signature

Declaration of your own contribution to the publications

Felix Nikolaus Wirth contributed to the following articles:

The research performed in the first part of my doctoral project was published in this paper: Wirth FN, Meurers T, Johns M, Prasser F. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. BMC Med Inform Decis Mak. 2021 Aug 12;21(1):242. doi: 10.1186/s12911-021-01602-x. PMID: 34384406.

Contribution:
Based on his expertise in the field, Felix Nikolaus Wirth conceptualized the systematization (i.e., method) and mapped existing approaches into the framework. Moreover, he analyzed existing approaches and collected and interpreted the data to form clusters of approaches sharing similarities (i.e., results). Finally, he drafted all sections of the manuscript and was responsible for the revisions.

The research performed in the second part of my doctoral project was published in this paper: Wirth FN, Kussel T, Müller A, Hamacher K, Prasser F. EasySMPC: a simple but powerful no-code tool for practical secure multiparty computation. BMC Bioinformatics. 2022 Dec 9;23(1):531. doi: 10.1186/s12859-022-05044-8. PMID: 36494612.

Contribution:
Felix Nikolaus Wirth conceptualized the method and designed and implemented about 90% of the graphical software application, the message exchange method using e-mail and the command-line interface (i.e., methods and results). Moreover, he designed and performed the experiments for the performance analysis and evaluated and interpreted the results. Finally, he drafted the entire manuscript and was responsible for the revisions.

_____

Signature, date and stamp of first supervising university professor / lecturer

_____

Signature of doctoral candidate

Printing copy of the first publication

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Privacy-preserving data sharing infrastructures for medical research: systematization and comparison

Felix Nikolaus Wirth*, Thierry Meurers, Marco Johns and Fabian Prasser

Abstract

Background: Data sharing is considered a crucial part of modern medical research. Unfortunately, despite its advantages, it often faces obstacles, especially data privacy challenges. As a result, various approaches and infrastructures have been developed that aim to ensure that patients and research participants remain anonymous when data is shared. However, privacy protection typically comes at a cost, e.g. restrictions regarding the types of analyses that can be performed on shared data. What is lacking is a systematization making the trade-offs taken by different approaches transparent. The aim of the work described in this paper was to develop a systematization for the degree of privacy protection provided and the trade-offs taken by different data sharing methods. Based on this contribution, we categorized popular data sharing approaches and identified research gaps by analyzing combinations of promising properties and features that are not yet supported by existing approaches.

Methods: The systematization consists of different axes. Three axes relate to privacy protection aspects and were adopted from the popular Five Safes Framework: (1) safe data, addressing privacy at the input level, (2) safe settings, addressing privacy during shared processing, and (3) safe outputs, addressing privacy protection of analysis results. Three additional axes address the usefulness of approaches: (4) support for de-duplication, to enable the reconciliation of data belonging to the same individuals, (5) flexibility, to be able to adapt to different data analysis requirements, and (6) scalability, to maintain performance with increasing complexity of shared data or common analysis processes.

Results: Using the systematization, we identified three different categories of approaches: distributed data analyses, which exchange anonymous aggregated data, secure multi-party computation protocols, which exchange encrypted data, and data enclaves, which store pooled individual-level data in secure environments for access for analysis purposes. We identified important research gaps, including a lack of approaches enabling the de-duplication of horizontally distributed data or providing a high degree of flexibility.

Conclusions: There are fundamental differences between different data sharing approaches and several gaps in their functionality that may be interesting to investigate in future work. Our systematization can make the properties of privacy-preserving data sharing infrastructures more transparent and support decision makers and regulatory authorities with a better understanding of the trade-offs taken.

Keywords: Biomedical data sharing, Privacy, Usefulness, Systematization, Distributed computing, Secure multi-party computing, Data enclave

*Correspondence: felix-nikolaus.wirth@charite.de
Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany

© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Introduction

Data sharing is the practice of making data from research and healthcare available for secondary purposes and to third parties. This enables data-driven medical research, which promises to significantly improve public health as well as prevention, diagnosis, treatment and follow-up care [1, 2]. It is advocated at both national and international levels [3–5] and is steadily becoming a standard practice in biomedical research [6]. The benefits of data sharing include larger sample sizes and the ability to generate new insights and to replicate results in times of increasing personalization of medicine. Data sharing is also associated with higher citation rates [7, 8] and promoted by several funding agencies [9, 10].

Despite its promises, several obstacles make data sharing difficult and often even impossible. An important obstacle are legal issues [11], caused by severe restrictions on the processing of personal medical data imposed by national and international data protection laws. Important examples include the US Health Insurance Portability and Accountability Act (HIPAA) [12] and the EU General Data Protection Regulation (GDPR) [13]. To process data in compliance with these regulations, organizational and legal procedures need to be implemented to protect the privacy of patients and research participants. An important prerequisite for data processing in medical research is usually informed consent. However, collecting consent can be difficult and is not always feasible [14], in particular when data is to be shared retrospectively at large scale. An alternative that is often suggested (and permitted in many jurisdictions) is anonymization, i.e. the altering of data in such a way that individual patients and research participants cannot be identified, rendering the data non-personal [15]. However, a trade-off between privacy protection and the quality and hence utility of output data needs to be considered in this process [16]. In this context, the complexity and heterogeneity of clinical and research data makes effective anonymization without disproportionately negative effects on data quality sometimes difficult and in some cases even impossible [17]. How strict the requirements for anonymization are depends on the applicable legislation. For example, while the HIPPA Privacy Rule [12] provides an interpretable and implementable framework, anonymization under the GDPR is more difficult due to a lack of concrete requirements and resulting heterogeneous policies and legal interpretations [18]. In addition, researchers often do not want to lose control of their data and institutions are often reluctant to disclose data that is considered confidential from a business perspective, e.g. for competitiveness reasons [19].

These challenges can be tackled by implementing infrastructures that enable analyzing data stored in distributed databases and computing a common result without exchanging individual-level data [10]. In the context of this work, we refer to such methods as "data sharing infrastructures", which involve different parties or sites (e.g. hospitals) in a joint analysis.

On the methodological side, there are different options for implementing this process. One well-known example is the exchange of aggregated statistics (see e.g. [20]), which are then combined to a common result, comparable to a meta-analysis. Another example is cryptographic protocols (e.g. [21]), enabling different parties to jointly process a function on their private data without revealing each other's input. Such modern secure multiparty computing schemes often employ homomorphic encryption, which supports operations such as addition and multiplication on encrypted data [22].

Technology infrastructures built on these approaches have already been successfully used to investigate a range of medical questions. Examples include studies of associations of maternal movement and newborn birth size [23], outcomes of partial or full knee replacement [24], treatment patterns for comorbidities of patients suffering from cancer [25], survival of patients with intrahepatic cholangiocarcinoma [26] and of interactions between food intake as well as gut bacteria and metabolite patterns [27]. Other projects have implemented manual processes for distributed data analysis, such as the 4CE consortium [28], which focuses on the clinical trajectory of COVID-19 patients or a study carried out in the German Medical Informatics Initiative, focusing on multimorbidity and rare diseases [29].

Objectives and contributions

Despite the fact that privacy-preserving infrastructures are often considered to be the most important enabler for comprehensive data sharing in the medical domain and despite the multitude of technological approaches available and studies that have successfully utilized such technologies (see above), these infrastructures are only rarely used for sharing healthcare and medical research data today. We believe that one of the main reasons for this is uncertainties for decision makers and regulatory authorities regarding the exact characteristics of such infrastructures, particularly regarding the degree of privacy protection and anonymity for data subjects they provide. Indeed, as we will show in this article, there are fundamental differences between current solutions.

As a first step towards making the properties of data sharing infrastructures more transparent, the aim of this work is to introduce a systematization of general techniques and their properties along two dimensions. Firstly,

the systematization is intended to structure the design space, as a development step towards tools for comprehensively assessing the privacy protection properties of data sharing infrastructures. Secondly, we also believe that the systematization can contribute to developing instruments for assessing the usefulness of data sharing infrastructures, i.e. the impact that their protection mechanisms have on options to analyze data compared to the simple (but often not feasible) approach of pooling all data in a common database.

The need for a framework for comparing different approaches to data sharing is also illustrated by the fact that several previous papers have been published on related topics (see section "Comparison with prior work"). However, our work is fundamentally different in that we do not only consider specific types of solutions (e.g., based on cryptographic methods) and aim at systematically mapping the usefulness dimension in addition to the privacy protection dimension. This comes at the expense of a higher degree of abstraction.

To show that our approach is practicable, we used it to perform a high-level analysis and comparison of several existing solutions. In summary, our work provides the following contributions:

(1) We present a high-level and technology-agnostic framework consisting of three axes describing the degree of protection and three axes describing the degree of usefulness provided by data sharing infrastructures.
(2) We use this framework to analyze and compare ten different real-world data sharing platforms. Our results show that they can be grouped into three general types of solutions with common properties.
(3) Based on our results we derive insights into research gaps that may be worthwhile to investigate when developing next-generation data sharing infrastructures.

Methods

Trade-off between privacy protection and usefulness

Data sharing would be easy to implement if all relevant data could simply flow freely and be stored in a common database. As mentioned above, this is not possible in practice, however. Any attempt to take measures to meet privacy protection requirements inevitably leads to limitations in comparison to this basic approach. These limitations may relate, for example, to the time that the data sharing process takes or to the number of analysis methods supported. This fundamental conflict between unrestricted processing of data and the protection of the privacy of data subjects is well known in the field of privacy-enhancing technologies. An important example is data anonymization, where, as also mentioned above, the quality of output data often must be traded off against the degree of privacy protection achieved (see e.g. [30]).

Similar trade-offs must be made when designing and implementing privacy-preserving data sharing infrastructures. Figure 1 provides an abstract, schematic illustration of this trade-off. It is derived from the concept of risk-utility curves, as used in data anonymization research (see e.g. [31]). The y-axis describes the level of privacy protection, while the x-axis describes the level of usefulness of an infrastructure. Examples of aspects that could be captured by the x-axis include the spectrum of functionalities offered, how scalable their implementations are and how much work is required to add new functionalities.

There are two extreme types of approaches. Approaches located in the top-left corner significantly limit the amount of data shared, e.g. only patient or research participant counts, which typically implies a very high degree of protection. Approaches located in the bottom-left corner exchange fine-grained data in nearly unmodified form, e.g. by pooling all data in a central database which is open for access by researchers. Obviously, this would be extremely useful, but offers little privacy protection.

In between these two extremes, there is a broad spectrum of potential solutions based on different trade-offs between privacy protection and usefulness. To be relevant, those data sharing approaches need to provide added value in comparison to the basic approaches, i.e. they need to significantly reduce privacy risks, while

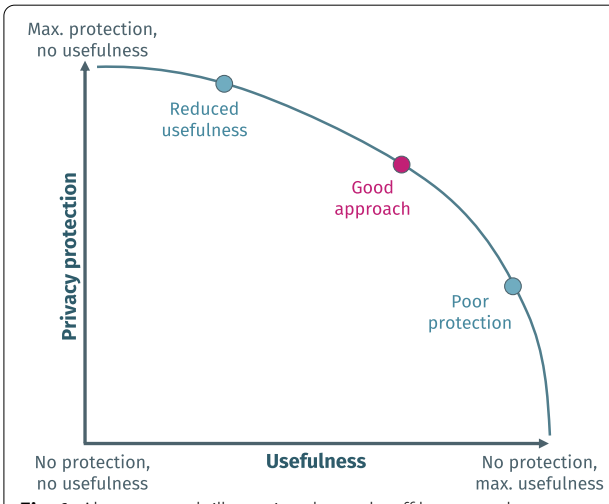


Fig. 1 Abstract graph illustrating the trade-off between the degree of privacy protection and the usefulness of a data sharing infrastructure

maintaining a high degree of usefulness. In the graph, this is indicated by the non-linear relationship between the extreme points.

One example is the aforementioned meta-analysis approach in which more than counts can be exchanged when appropriate safeguards are implemented (e.g. for regression coefficients [32]). Still, functionality is limited, as only aggregated data from individual sites can be included in the analysis, hence reducing the number of (scientific) questions that can be answered. At the same time, privacy is relatively easy to protect by making sure that the aggregate data released does not leak sensitive personal information.

A framework for systematizing properties of data sharing techniques

For assessing the degree of privacy protection and the usefulness provided by data sharing approaches, we propose a first systematization containing three axes for each of these aspects. These axes are illustrated in Fig. 2 and will be explained in more detail in this section.

Aspect 1: Assessing the degree of privacy protection provided

As a baseline for assessing the degree of protection provided we suggest to apply the Five Safes Framework, which was developed by Desai, Ritchie and Welpton as a general framework for reasoning about privacy protection when sharing data [33] (important examples are discussed in Section "Comparison with prior work").

The Five Safes Framework specifies five different axes, which are illustrated in Fig. 3: (1) Only *Safe People*, e.g. trustworthy researchers, should be provided with access

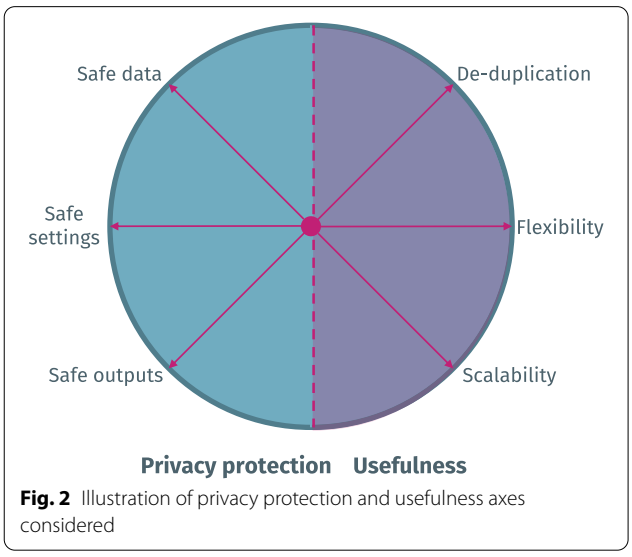


Fig. 2 Illustration of privacy protection and usefulness axes considered

Fig. 3 Elements of the five safes framework. Axes with relevance to this work are highlighted in blue

to data (cf. the British Office for National Statistics research and data access policy [34]), (2) only *Safe Projects* should be carried out, e.g., analyses that respect patient privacy and which are appropriate from an ethical perspective, (3) only *Safe Data* should be processed meaning that identifiability should be reduced to an acceptable minimum already on the level of input data (cf. the principle of data minimization under the GDPR and the Minimum Necessary Standard of HIPAA [10]), (4) *Safe Settings* should be used for providing access or performing analyses, which reduces the likelihood that sensitive data is leaked during processing and (5) *Safe Outputs* should be guaranteed (e.g., by ensuring that the output of analyses does not disclose sensitive personal information).

For our framework we will only consider the technical aspects of the Five Safes Framework and thus exclude the first two axes, *Safe People* and *Safe Projects*. The reason is that these aspects need to be either addressed on an organizational level (e.g. ethics committee / Institutional Review Board (IRB) approval) or with technical solutions that are not directly related to data sharing (e.g. Authentication and Authorization Infrastructures). In the context of data sharing, there are specific measures that can be taken along the remaining technical axes:

Axis 1.1: Safe data

Data provided as input to analyses supported by data sharing is considered safe if the identifiability of patients or research participants has been reduced. *Safe Data* can for example be obtained by anonymization, aggregation or encryption. Protection achieved with the first two techniques may be irreversible, while it may be possible to decrypt encrypted data at the end of the process. Even if anonymization or aggregation has limitations, residual risks of identifiability can potentially be managed by implementing safeguards along the other axes.

Axis 1.2: Safe settings

The setting in which distributed data is processed is considered safe if no or at least only some data is leaked during processing. A well-known example of a *Safe Setting* are virtual data access environments, in which data can be analyzed without handing out individual-level data, e.g. through a remote desktop connection. Infrastructures using cryptographic secure multi-party computing protocols also provide a secure setting in which data can

be analyzed in an encrypted form only and only mutually calculated results can be decrypted [35] (more details will be provided in the "Results" section). However, even with such safe settings being used to perform analyses, additional efforts may need to be made to ensure that the results are also safe.

Axis 1.3: Safe outputs

The result calculated using a data sharing infrastructure is considered safe, if the resulting data disclosed to the users of the infrastructure is non-identifiable/non-personal. One way of achieving this is to only allow computations producing aggregate data. However, this must be carefully designed, as e.g. disclosing statistical tables with small cell counts can reveal details about individuals [36]. To mitigate this risk, anonymization methods can be used to transform data before it is being disclosed. For example, data points can be rounded up, they can be omitted or random noise can be added [37]. A state-of-the-art technique to provide *Safe Outputs* is Differential Privacy which formulates a general mathematical property for data processing algorithms that, if parameterized correctly, renders output data non-identifiable [38]. We note that a data sharing infrastructure will automatically provide *Safe Outputs* when *Safe Data* is provided as input (cf. the meta-analysis approach).

Aspect 2: Assessing the usefulness of data sharing technologies

As a first step, we suggest to assess the usefulness of infrastructures for sharing medical data in terms of three different axes that reflect important requirements in multi-institutional medical research: (1) *Deduplication/record-linkage*, which refers to the ability to combine data from different sources while taking into account that some records might relate to one another (e.g. to the same patient), (2) *Flexibility*, which reflects the degree to which a solution is able to support different types of statistical analyses and use cases as well as adapt to different analytical requirements as they can change over time and (3) *Scalability*, that refers to how an infrastructure performs when the amount of data or the complexity of an analysis increases.

Axis 2.1: De-duplication/record-linkage

This axis is related to the ability to resolve different types of data distribution, which are sketched in Fig. 4. Most data sharing infrastructures are able to resolve horizontal distribution of data but ignore potential relationships on the level of individuals. This is for example the case with meta-analyses in which patient data from different hospitals is simply added to a larger sample without checking for population overlap. In order to determine or resolve such overlap, privacy-preserving methods for reconciling records belonging to the same individuals must be implemented, which is non-trivial. This becomes even more challenging, when also vertical distribution is to be resolved. A typical example is the need to integrate different types of data for the same patients stored at different locations (e.g. at hospitals and health insurances). Procedures allowing for such a cross-site duplicate resolution range from probabilistic linkage algorithms [39] and cross-site pseudonymization methods to secure linkage based on encrypted identifying information using secure multi-party computing protocols [40, 41]. This results in different characteristics with regard to risks and usefulness, which manifests itself, for example, in the possibility of verifying the correctness of linkage results. A

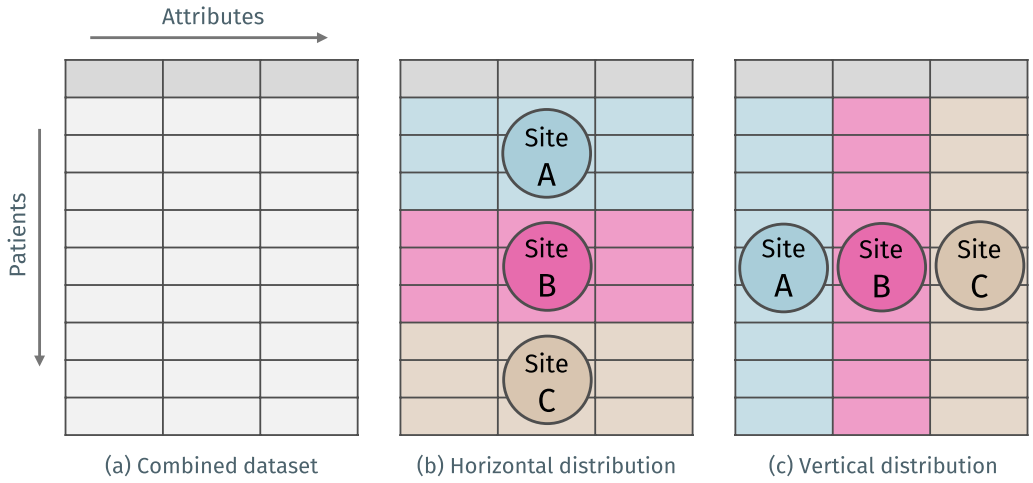


Fig. 4 Horizontal and vertical data distribution

cross-site pseudonymization procedure poses the greatest risks but provides the highest linkage quality, whereas probabilistic linkage and cryptographic methods offer a very high level of protection, but make it difficult to verify the results. The associated risks are reflected by axes 1.1, 1.2 and 1.3, while the usefulness of de-duplication and record-linkage is reflected by this axis. For the sake of clarity, we will simply refer to this axis as "*De-duplication*" in the remainder of this article.

Axis 2.2: Flexibility

This axis refers to the ability of infrastructures to support a range of different analyses and to its extensibility to future use cases. For example, some of the solutions analyzed in this article have been tailored towards a limited set of very specific functionalities (e.g. cohort selection). On the other hand, some solutions are based on generic frameworks that provide a high degree of extensibility and options to integrate new analysis methods. In between these two extremes, there are solutions, e.g. based on meta-analyses, which offer a certain degree of flexibility, but only support some types of analyses. For example, the quality of survival analyses might be inconsistent, analyses of subgroups might require additional efforts for each subgroup and longitudinal studies as well as explorative investigations and assessments of data quality may be difficult to perform [42, 43]. There are also differences regarding the effort required to integrate new methods into different types of solutions. For example, integrating new types of analyses into solutions based on secure multi-party computing requires developing implementations using special cryptographic primitives, which is time-consuming and requires expert-level knowledge of cryptography.

Axis 2.3: Scalability

This axis refers to the ability of an infrastructure to function well, i.e. to return a result to the analysis performed within a reasonable timeframe and with a reasonable demand for compute and storage resources, when load is increased (this is also called load scalability [44]). Within the context of data sharing infrastructures, an increase in load can be caused by an increase in the volume (e.g. number of patients) or dimensionality (e.g. number of attributes per patient) of the data analyzed, or the number of sites participating in the sharing process. *Scalability* is a particular challenge for approaches based on secure multiparty computing, as current state-of-the-art approaches are known to not scale well with respect to both of these aspects. It general, it can be said that the performance of all secure multiparty computing methods is determined by the number of messages exchanged between the parties involved, the required number of rounds of communication between the parties and the computational overhead per round. It should be noted, however, that the exact increase in computational complexity depends on the particular type of method used [45]and the operation performed [46].

Results

In this section, we present the results of an application of the framework proposed for an analysis of a range of well-known data sharing infrastructures for medical data that exhibit different characteristics along the axes suggested. We note that some infrastructures are relatively generic and can be used to implement different methods with different characteristics. In these cases, we analyzed typical applications of the infrastructures and present alternative use cases in the "Discussion" section. In particular, we analyzed the following solutions: SHRINE/i2b2 [47], dataSHIELD [20], OMOP/OHDSI [48], Personal Health Train [49], Clinerion Patient Network Explorer [50], TriNetX [51], MedCo [52], Sharemind MPC [53] and examples implementing the popular Data Enclave concept [54, 55]. Based on common privacy protection properties of the approaches studied, we assigned them to three different categories: (1) distributed data analysis, (2) cryptographic secure multi-party computing approaches and (3) data enclaves.

Distributed data analysis

One category, termed distributed data analysis, contains approaches that exchange aggregated and potentially anonymized data only. This non-personal data is generated locally at the participating sites and then merged across locations using meta-analysis methods. Hence, regarding our framework, only aggregated or anonymized data (and thus *Safe Data*) is exchanged (axis 1.1), no *Safe Setting* is hence needed (axis 1.2) and *Safe Outputs* are provided by design (axis 1.3). However, there are significant limitations regarding the analytical utility of these types of data sharing approaches. None of the solutions analyzed from this category supports *De-duplication* (axis 2.1), since vertical integration can only be conducted with additional measures (see "Discussion" section). Moreover, some of the approaches in this category are very specific and others are quite generic (*Flexibility*, axis 2.2), while all share the disadvantages of meta-analyses described in section "Aspect 2: Assessing the usefulness of data sharing technologies", such as limited possibilities to perform subgroup analyses. However, all approaches provide a high degree of computational *Scalability*, as computations can be offloaded to the participating sites effectively (axis 2.3). Important examples of approaches in this category are:

- *SHRINE/i2b2* Informatics for Integrating Biology & the Bedside (i2b2) is an open-source clinical data warehouse used in various projects worldwide [47]. The Shared Health Research Information Network (SHRINE) is an extension of i2b2 for distributed analysis [56]. It allows the creation of a network of peer sites, in which aggregated results of queries are collected. SHRINE is for example used in a registry of pediatric patients with rheumatic disease [57] or in a network supporting clinical trial recruitment [58]. The solution is specific, since it has been designed specifically to support cohort selection functions (*Flexibility*, axis 2.2).
- *DataSHIELD* This software supports distributed analyses based on the R statistical computing environment [20]. It creates a network of server nodes that connect to local instances of R. Through a client node, researchers can then send commands which are distributed to the local sites to calculate aggregated results without individual-level data leaving the sites. DataSHIELD has been deployed, for instance, in a network that investigates interactions of ageing, mental well-being and environment [59] and it is a generic solution, since it supports a range of analysis methods based on R (*Flexibility*, axis 2.2).
- *OMOP/OHDSI* The Observational Health Data Sciences and Informatics (OHDSI) project [48] has developed the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), which can be used to create highly structured and standardized local databases for real-world evidence studies. For distributed analyses, scripts can be executed at the sites to derive aggregate data that can then be combined in meta-analyses. This process is also supported by a range of tools provided by the OHDSI community. In practice, this approach has for example been utilized in a study of models for predicting stroke in women [60]. The European Health Data Evidence Network (EHDEN) [61] aims to foster the adoption of OMOP/OHDSI in Europe. The approach is generic, since a wide range of analyses is supported (*Flexibility*, axis 2.2).
- *Personal Health Train* The Personal Health Train (PHT) is data sharing concept developed by different private and public contributors [49]. It is based on a train analogy: (1) the data sources are called train stations, (2) the data analysis methods (e.g. query and merge procedures) are called trains. In all current implementations only aggregated data leave the stations towards the trains, hence implementing a meta-analysis approach. The PHT has for instance been used to realize a study on distributed learning for predicting the post-treatment two-year survival of lung cancer patients [62]. It is a generic solution conceptualizing a container-based data sharing infrastructure that can be used to implement wide a range of meta-analysis approaches (*Flexibility*, axis 2.2).
- *Clinerion Patient Network Explorer* and *TriNetX* Both the Patient Network Explorer by Clinerion [50] and the software by TriNetX [51] are parts of propriety data sharing networks for hospitals established by these companies. After installing the software, local nodes in the hospitals provide interfaces for central services to collect aggregated data, for instance the number of patients meeting certain inclusion criteria. As an example, TriNetX has been used to collect data for investigating the risk of COVID-19 for people suffering from intellectual and developmental disabilities [63]. Both solutions can be described as specific, as privacy protection is implemented by restricting the analysis methods supported (*Flexibility*, axis 2.2).

Secure multi-party computation

Approaches using cryptography-based secure multi-party computation protocols to ensure that only encrypted individual-level data leaves the participating sites form an important additional category of data sharing infrastructures. Typically, it is also ensured that only analytical results aggregating the data from multiple sites can be decrypted at the end of a computation (thus also providing protection on the institutional level). As a result, only *Safe Data* (i.e. encrypted data) is exchanged (axis 1.1) in a *Safe Setting*, as data is not disclosed during processing (axis 1.2). All solutions identified that fall into this category further implement specific analysis methods that ensure that only *Safe Outputs* are disclosed (axis 1.3). We note, however, that this is not an inherent property of cryptographic approaches but a result of performing secure analyses or perturbing output data by the approaches investigated. It is well-known that *Scalability* can be a problem for secure multi-party computation protocols (axis 2.3). Performance is often non-linear in the number of participating sites, implementations require a lot of computational resources and low-latency network connections with a high transmission rate, which can typically not be provided when data is shared over the internet. Whether or not duplicates can be detected and resolved (*De-duplication*, axis 2.1) and different types of analyses can be performed (*Flexibility*, axis 2.2) depends on the exact implementation. Important examples of approaches from this category are:

- *MedCo* The open source software MedCo uses additively homomorphic encryption to enable research-

ers to perform analyses on encrypted data across sites [52]. The analysis results are encrypted and can only be decrypted by authorized investigators. MedCo is implemented as an extension to i2b2 (analogously to SHRINE). The software, for example, forms the backbone of the SCOR network for sharing data on patients with COVID-19 [64]. The software focuses on cohort exploration and survival analysis. MedCo does not support resolving duplicates (*De-duplication*, axis 2.1) and is specific, as it only supports a limited set of functionalities and extensions require implementations to be developed based on the cryptographic methods used by the software (*Flexibility*, axis 2.2).

- *Sharemind MPC* This proprietary software has been developed by the company Sharemind. Similar to MedCo it enables computations on encrypted data hosted at multiple sites without decrypting it first [53]. The software is oriented towards data scientists. Analyses can either be designed in a proprietary programming language or in an environment which resembles the R statistics programming environment. The solution has, for example, been used to analyze 10 million synthetic health records distributed to 1,000 health centers that also involved detecting and removing duplicates (*De-duplication*, axis 2.1) [65]. Sharemind MPC provides a generic framework for privacy-preserving data sharing (*Flexibility*, axis 2.2).

Data enclaves

The third category of approaches consists of implementations of the data enclave concept, in which individual-level data of one or multiple sites is submitted to a data custodian maintaining a secure environment for data access [66]. Eligible researchers can run queries against the data stored by the custodian, the results of which are checked for anonymity before they are returned. Hence, individual-level, non-safe data is exchanged (*Safe Data*, axis 1.1) but access is restricted through a *Safe Setting* (axis 1.2) which ensures that no data is leaked and that output data is safe (*Safe Outputs*, axis 1.3). On the usefulness dimension, duplicate resolution is supported (*De-duplication*, axis 2.1) and large datasets as well as data from many participant sites can be shared in scalable manner (*Scalability*, axis 2.3). However, real-world implementations differ regarding their extensibility and *Flexibility* (axis 2.2). Important examples of data enclaves are:

- *Scottish National Safe Haven* This enclave is operated by the Scottish National Health Services (NHS) and provides access to various health datasets [54]. Data is stored in pseudonymized form to enable record linkage. Data access is provided through a virtual network with no internet access and no ability to install custom software. The infrastructure has, for example, been used to study temporal trends in breast cancer incidence [67]. The solution is somewhat generic, as typical data analysis methods are supported, but extensibility is limited as additional software, packages and functionalities can only be implemented by the enclave (*Flexibility*, axis 2.2).
- *US Center for Medicare and Medicaid Services Virtual Research Data Center* This enclave is operated by the US Center for Medicare and Medicaid Services and provides access to claims data combined with other types of medical data [55]. To ensure that output data is safe, researchers are only allowed to export aggregated information which is reviewed and screened for identifiability before it can be downloaded [68]. The system has, for example, been used for a study on the relative risk of Alzheimer's disease among patients with prostate cancer who received androgen deprivation therapy [69]. The solution is specific, since its software and functionalities focus on integration and analysis of claims data (*Flexibility*, axis 2.2).

Discussion

Principal results

In the previous sections, we have proposed a schema for systematizing privacy-preserving data sharing infrastructures for medical research. We applied this framework to study a wide range of solutions proposed and found that they can be assigned to three distinct categories, based on common properties. Table 1 summarizes the results of our analysis.

As can be seen from this summary, most solutions identified fall into the category of distributed data analyses. One reason for this could be the fact that the technical complexity of this approach is relatively low, while it supports a fairly wide range of use cases. In comparison, secure multi-party computation is quite complex from a technical perspective and data enclaves are difficult to set up in some legislations, as individual-level data may not be allowed to leave the institutions in which it was initially collected. Distributed data analysis, however, reaches its limits when analyses on individual-level are needed or complex record-linkage and duplicate detection functionalities are required. Secure multi-party computation and data enclaves are relatively new approaches to medical data sharing, which can provide more functionalities. For them to be used even

Table 1 Results of our analysis of solutions for privacy-preserving data sharing

Approach	Year of publication	Category	1. Privacy protection			2. Usefulness		
			1. Safe data	2. Safe settings	3. Safe outputs	1. De-duplication	2. Flexibility	3. Scalability
SHRINE/i2b2	2008	Distributed data analysis	Yes	No	Yes[b]	No	Specific	Yes
dataSHIELD	2010	Distributed data analysis	Yes	No	Yes[b]	No	Generic	Yes
OHDSI	2014	Distributed data analysis	Yes	No	Yes[b]	No	Generic	Yes
Personal Health Train	2017	Distributed data analysis	Yes	No	Yes[b]	No	Generic	Yes
Clinerion	2015	Distributed data analysis	Yes	No	Yes[b]	No	Specific	Yes
TriNetX	2015	Distributed data analysis	Yes	No	Yes[b]	No	Specific	Yes
MedCo	2018	Secure multi-party computation	Yes[a]	Yes	Yes	No	Specific	No
ShareMIND	2008	Secure multi-party computation	Yes[a]	Yes	Yes	Yes	Generic	No
Scottish National Safe Haven	2015	Data enclave	No	Yes	Yes	Yes	Generic	Yes
Virtual Research Data Center	2014	Data enclave	No	Yes	Yes	Yes	Specific	Yes

[a] The processed data is encrypted individual-level data and thus safe

[b] *Safe Outputs* is an implicit result of providing *Safe Data* as input

more widely, technical challenges (e.g. regarding suitable cryptographic protocols) as well as legal challenges (e.g. regarding the question whether encrypted data be considered non-personal or what an appropriate legal status for data custodians could look like) will need to be overcome. To accelerate work on these issues, policymakers should consider incentives for making innovative choices regarding data sharing architectures.

Comparison with prior work

Our work builds on the Five Safes framework to systematize privacy protection. In prior work, the framework has already been used to study data sharing in official statistics [70], social and political sciences [71] and psychology [72]. In the biomedical domain, the framework has been adopted to model risk-based anonymization approaches [73]. To the best of our knowledge, our work is the first to apply the framework to common biomedical data sharing infrastructures, however. Moreover, we have complemented the Five Safes framework for modeling privacy protection with additional axes for systematizing the usefulness of data sharing technologies, considering common requirements from biomedical research. Other articles analyzing data sharing infrastructures, such as the work by Foster [71], are not systematic and do not focus on biomedical research.

Other frameworks for data sharing in biomedical research have been proposed, which can also be used to analyze different technical approaches. These focus on other aspects, however. For example, Knoppers [74] proposed a framework for the sharing of genomic data with a particular emphasis on trust, responsible research and oversight using organizational and legal safeguards. This is comparable to the non-technical axes *Safe People* and *Safe Projects* of the Five Safes Framework [33]. Moreover, Aziz et al. [75] presented an overview of privacy-preserving techniques for sharing genomic data, which is particularly sensitive and difficult to protect from privacy breaches. Hence, the paper puts a specific focus on cryptographic methods tailored towards genomic data sharing, which provide strong and provable degrees of protection. Compared to our approach their framework used for comparisons is rather specific, focusing on cryptographic algorithms and their technical properties and less on off-the-shelf, more generic infrastructures. Still, many of the aspects used by Aziz et al. in their comparisons are partially congruent to aspects of our framework (e.g. execution time, memory usage and network communication as aspects of *Scalability*, secure

computations and output privacy as synonyms for *Safe Settings* and *Safe Outputs*, and accuracy as an aspect of *Usefulness*), which can be seen as an additional indicator for the broad applicability of our framework. Also Mittos et al. [76] presented a systematization of privacy-enhancing technologies for processing genomic data. However, their work focuses on many different types of processing, from which data sharing is just one example. Still, many of the open issues identified, such as the computational costs of some approaches and the need to improve the usefulness of results are in-line with our findings. Naveen et al. [77] presented an overview of applications, challenges and solutions for genomic data processing, which also includes aspects of data sharing. Their work contains lists of known privacy threats and specific approaches for implementing different use cases while mitigating those threats. Along these lines they systematically analyze open challenges within different application areas, but do not propose a common systematization spanning all of them. Notably, they also highlight some of the challenges mentioned in our work, such as the inherent trade-off between degrees of protection and usefulness. Thapa et al. [78] presented an overview of data sharing technologies for the more general area of "precision health", also focusing primarily on cryptographic methods and methods requiring specific hardware support (e.g. Trusted Computing Environments). Consequently, the aspects used in their comparison of different approaches are quite similar to the aspects used by Aziz et al., which are well aligned with our more high-level framework as discussed above. In addition to that, they analyzed specific applications of data sharing frameworks, e.g. for distributed machine learning. The axes used for comparing such solutions could serve as a basis for future extensions of our framework (see section "Limitations, future work and open research questions").

A framework for real-world multi-database studies has been presented by Toh [79]. On a conceptual level, this framework is most closely related to our work. However, it puts a strong focus on study design and feasibility and thus only considers weighing analytic flexibility with privacy protection on the utility and risk axes as well as trading off data pooling and distributed analyses on the technology axes. Finally, a comprehensive, yet unsystematic, overview of infrastructures for sharing data on COVID-19 has been presented by Raisaro et al. [64].

Limitations, future work and open research questions

We note that the systematization proposed is abstract and of a qualitative nature. It is hence only suited for performing initial high-level comparisons of different solutions in the field as exemplified by the results of our analysis of selected implementations. Although a rigorous and formal framework would be desirable to enable more detailed comparisons, constructing such a framework is highly challenging. Important reasons can be found in a recent comment by Richie and Green [80] in which the authors advocate for the qualitative nature of the Five Safes framework. Aziz et al. [75] also report challenges in identifying technical and quantitative criteria that are general enough to apply to different types of approaches and that at the same time can be used for specific comparisons.

At a more fundamental level, even the quantitative modeling of privacy risks and usefulness is still an open research problem. Both aspects can only be captured by models that make very specific assumptions, which in turn may not apply to all projects and usage scenarios. For example, a recent overview by Wagner and Eckhoff lists 80 different formal privacy models [81]. However, some data sharing infrastructures and approaches support different privacy models to provide *Safe Data* and *Safe Outputs*, e.g. Differential Privacy [38] or solutions limiting the uniqueness of disclosed data, such as cell suppression [82] or k-anonymity [83]. In future work, we plan to extend our framework by incorporating the most common models. Regarding the usefulness of solutions, some of the more fine-grained axes used in [75, 78] might serve as a starting point. One example is *Accuracy*, which reflects the impact of privacy models on output data quality and hence captures the risk-utility trade-off inherent to such technologies.

The results of our analysis of the current landscape of solutions can also provide insights into potential directions for future work on data sharing methods. One important example is the low number of solutions supporting de-duplication or record linkage. When analyzing horizontally distributed data, the inability to identify and resolve population overlap can significantly reduce the quality of results [84]. If a study intends to analyze vertically distributed data, record linkage is crucial, as different data sets need to be combined on a patient-level. One important example is research on rare diseases, as patients with such conditions typically visit a wide range of healthcare providers and relevant data for each patient is therefore inherently distributed. Future work could be carried out to extend distributed data analysis infrastructures with record-linkage functionalities, e.g. by enriching data with secure record linkage tokens [85]. Also, secure multi-party computation environments could be extended with libraries including different record-linkage algorithms (see [86] for a recent example). Moreover, future work could explore ways to provide strong protection guarantees for inherently flexible approaches, such as the Personal Health Train. This could, for example, be achieved by integrating libraries

providing support for a wide range of privacy-preserving analysis functions within such infrastructures. Finally, a challenge with privacy-preserving data sharing infrastructures is that access to individual-level data in some cases cannot be provided at all, although access to data from at least one site is often needed to develop analysis algorithms that can then be executed in the distributed network. One approach to overcome this limitation is to provide synthetic data derived from the original data for this preparatory process (see [87] for a recent example in the context of distributed data analysis).

Conclusion

In this article, we proposed a high-level framework for analyzing and comparing privacy-preserving data sharing infrastructures for medical research. We believe that our framework makes the properties of data sharing approaches more transparent and can serve as a starting point for developing more comprehensive systematizations, ultimately supporting decision makers and regulatory authorities in gaining a better understanding of the trade-offs taken. We have shown that our systematization is of value, by using it to analyze existing solutions, showing that there are fundamental differences between them. Finally, our results also provide insights into gaps, regarding the systematization itself as well as the current landscape of data sharing infrastructures, that may be worth exploring in the future.

Authors' contributions

FW conceptualized the systematization, collected and interpreted the methods' data and drafted the manuscript. TM and MJ collected and interpreted the methods' data and reviewed the manuscript. FP supervised the work, conceptualized the systematization and reviewed the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

All data generated or analyzed during this study are included in this published article.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 7 January 2021 Accepted: 31 July 2021
Published online: 12 August 2021

References

1. Packer M. Data sharing in medical research. BMJ. 2018;360: k510. https://doi.org/10.1136/bmj.k510.
2. Weitzman ER, Kaci L, Mandl KD. Sharing medical data for health research: the early personal health record experience. J Med Internet Res. 2010. https://doi.org/10.2196/jmir.1356.
3. Carr D, Littler K. Sharing research data to improve public health. J Empir Res Hum Res Ethics. 2015;10:314–6. https://doi.org/10.1177/1556264615593485.
4. Pilat D, Fukasaku Y. OECD principles and guidelines for access to research data from public funding. Data Sci J. 2007;6:OD4–11. https://doi.org/10.2481/dsj.6.OD4.
5. Taichman DB, Backus J, Baethge C, Bauchner H, de Leeuw PW, Drazen JM, et al. Sharing clinical trial data—a proposal from the international committee of medical journal editors. N Engl J Med. 2016;374:384–6. https://doi.org/10.1056/NEJMe1515172.
6. Krumholz HM. Why data sharing should be the expected norm. BMJ. 2015. https://doi.org/10.1136/bmj.h599.
7. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. PLoS ONE. 2007;2:e308.
8. Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. PeerJ. 2013;1:e175. https://doi.org/10.7717/peerj.175.
9. Institute of Medicine. Sharing clinical research data: workshop summary. Washington, D.C: National Academies Press (US); 2013.
10. Hulsen T. Sharing is caring-data sharing initiatives in healthcare. Int J Environ Res Public Health. 2020. https://doi.org/10.3390/ijerph17093046.
11. Vis DJ, Lewin J, Liao RG, Mao M, Andre F, Ward RL, et al. Towards a global cancer knowledge network: dissecting the current international cancer genomic sequencing landscape. Ann Oncol. 2017;28:1145–51. https://doi.org/10.1093/annonc/mdx037.
12. Act A. Health insurance portability and accountability act of 1996. Public Law. 1996;104:191.
13. Regulation GDP. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. Off J Eur Union (OJ). 2016;59:294.
14. Williams G, Pigeot I. Consent and confidentiality in the light of recent demands for data sharing. Biom J. 2017;59:240–50. https://doi.org/10.1002/bimj.201500044.
15. Emam KE, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ. 2015. https://doi.org/10.1136/bmj.h1139.
16. Prasser F, Eicher J, Spengler H, Bild R, Kuhn KA. Flexible data anonymization using ARX—current status and challenges ahead. Softw Practice Exp. 2020;50:1277–304. https://doi.org/10.1002/spe.2812.
17. Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun. 2019;10:3069. https://doi.org/10.1038/s41467-019-10933-3.
18. Hansen J, Wilson P, Verhoeven E, Kroneman M, Kirwan M, Verheij R, et al. Assessment of the EU Member States' rules on health data in the light of GDPR. Brussels: EU publications; 2021. https://doi.org/10.2818/546193.
19. Ward MJ, Marsolo KA, Froehle CM. Applications of business analytics in healthcare. Bus Horiz. 2014;57:571–82. https://doi.org/10.1016/j.bushor.2014.06.003.
20. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014;43:1929–44. https://doi.org/10.1093/ije/dyu188.
21. Shi H, Jiang C, Dai W, Jiang X, Tang Y, Ohno-Machado L, et al. Secure Multi-pArty computation grid logistic regression (SMAC-GLORE). BMC Med Inform Decis Mak. 2016;16:89. https://doi.org/10.1186/s12911-016-0316-1.
22. Armknecht F, Boyd C, Carr C, Gjøsteen K, Jäschke A, Reuter CA, Strand M. A guide to fully homomorphic encryption. IACR Cryptol. ePrint Arch. 2015;2015:1192.
23. Pastorino S, Bishop T, Crozier SR, Granström C, Kordas K, Küpers LK, et al. Associations between maternal physical activity in early and late pregnancy and offspring birth size: remote federated individual level meta-analysis from eight cohort studies. BJOG Int J Obstetr Gynaecol. 2019;126:459–70. https://doi.org/10.1111/1471-0528.15476.
24. Burn E, Weaver J, Morales D, Prats-Uribe A, Delmestri A, Strauss VY, et al. Opioid use, postoperative complications, and implant survival after

unicompartmental versus total knee replacement: a population-based network study. Lancet Rheumatol. 2019;1:e229–36. https://doi.org/10.1016/S2665-9913(19)30075-X.
25. Chen R, Ryan P, Natarajan K, Falconer T, Crew KD, Reich CG, et al. Treatment patterns for chronic comorbid conditions in patients with cancer using a large-scale observational data network. JCO Clin Cancer Inform. 2020;4:171–83.
26. Hong N, Zhang N, Wu H, Lu S, Yu Y, Hou L, et al. Preliminary exploration of survival analysis using the OHDSI common data model: a case study of intrahepatic cholangiocarcinoma. BMC Med Inform Decis Mak. 2018;18:81–8. https://doi.org/10.1186/s12911-018-0686-7.
27. Oluwagbemigun K, Foerster J, Watkins C, Fouhy F, Stanton C, Bergmann MM, et al. Dietary patterns are associated with serum metabolite patterns and their association is influenced by gut bacteria among older German adults. J Nutr. 2020;150:149–58. https://doi.org/10.1093/jn/nxz194.
28. Brat GA, Weber GM, Gehlenborg N, Avillach P, Palmer NP, Chiovato L, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. Npj Digital Med. 2020;3:1–9. https://doi.org/10.1038/s41746-020-00308-0.
29. Kamdje-Wabo G, Gradinger T, Löbe M, Lodahl R, Seuchter SA, Sax U, et al. Towards structured data quality assessment in the German medical informatics initiative: initial approach in the MII demonstrator study. Stud Health Technol Inform. 2019;264:1508–9.
30. Li T, Li N. On the tradeoff between privacy and utility in data publishing. In: Elder J, Soulié Fogelman F, editors. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2009; Paris. New York: Association for Computing Machinery; 2009. p. 517–26. https://doi.org/10.1145/1557019.1557079.
31. Spengler H, Prasser F. Protecting biomedical data against attribute disclosure. Stud Health Technol Inform. 2019;267:207–14. https://doi.org/10.3233/SHTI190829.
32. Ritchie F. Disclosure control for regression outputs. WISERD data resources. 2011. https://wiserd.ac.uk/sites/default/files/documents//WISERD_WDR_005.pdf. Accessed 14 June 2021.
33. Desai T, Ritchie F, Welpton R. Five safes: designing data access for research. Bristol Business School Working Papers in Economics. 2016. https://www2.uwe.ac.uk/faculties/bbs/Documents/1601.pdf. Accessed 14 June 2021.
34. Office for National Statistics. ONS research and data access policy. n.d. https://www.ons.gov.uk/file?uri=/aboutus/transparencyandgovernance/datastrategy/datapolicies/onsresearchanddataaccesspolicy/attachmentresearchanddataaccesspolicy.pdf. Accessed 14 June 2021.
35. Evans D, Kolesnikov V, Rosulek M. A pragmatic introduction to secure multi-party computation. Foundations and Trends® in Privacy and Security. 2017;2(2-3). https://doi.org/10.1561/3300000019.
36. Murphy SN, Chueh HC. A security architecture for query tools used to access large biomedical databases. In: Kohane IS, editor. Proceedings of the AMIA Symposium; 2002; San Antonio. Philadelphia: Hanley & Belfus; 2003. p. 552–6 .
37. Bakken DE, Rarameswaran R, Blough DM, Franz AA, Palmer TJ. Data obfuscation: anonymity and desensitization of usable data sets. IEEE Secur Privacy. 2004;2:34–41. https://doi.org/10.1109/MSP.2004.97.
38. Dwork C. Differential privacy: a survey of results. In: Agrawal M, Du D, Duan Z, Li A, editors. Theory and Applications of Models of Computation. Proceedings of the 5th International Conference on Theory and Applications of Models of Computation; 2008; Xi'an. Berlin: Springer; 2008, p. 1–19. https://doi.org/10.1007/978-3-540-79228-4_1.
39. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. Int J Epidemiol. 2016;45:954–64. https://doi.org/10.1093/ije/dyv322.
40. Domadiya N, Rao UP. Privacy preserving distributed association rule mining approach on vertically partitioned healthcare data. Procedia Comput Sci. 2019;148:303–12. https://doi.org/10.1016/j.procs.2019.01.023.
41. Yigzaw KY, Michalas A, Bellika JG. Secure and scalable deduplication of horizontally partitioned health data for privacy-preserving distributed statistical computation. BMC Med Inform Decis Mak. 2017;17:1. https://doi.org/10.1186/s12911-016-0389-x.
42. Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. BMC Med Res Methodol. 2005;5:14. https://doi.org/10.1186/1471-2288-5-14.
43. Jones EM, Sheehan NA, Masca N, Wallace SE, Murtagh MJ, Burton PR. DataSHIELD—shared individual-level analysis without sharing the data: a biostatistical perspective. Norsk Epidemiol. 2012. https://doi.org/10.5324/nje.v21i2.1499.
44. Bondi AB. Characteristics of scalability and their impact on performance. In: Woodside M, Gomaa H, Menasce D, editors. Proceedings of the 2nd International Workshop on Software and Performance; 2008; Ottawa. New York: Association for Computing Machinery; 2000. p. 195–203. https://doi.org/10.1145/350391.350432.
45. Saia J, Zamani M. Recent results in scalable multi-party computation. In: Italiano GF, Margaria-Steffen T, Pokorný J, Quisquater J-J, Wattenhofer R, editors. SOFSEM 2015. Proceedings of the 41st International Conference on Current Trends in Theory and Practice of Informatics; 2015; Pec pod Sněžkou. Berlin: Springer; 2015. p. 24–44. https://doi.org/10.1007/978-3-662-46078-8_3.
46. Volgushev N, Schwarzkopf M, Getchell B, Varia M, Lapets A, Bestavros A. Conclave: secure multi-party computation on big data. In: Fetzer C, editor. Proceedings of the 14th EuroSys conference; 2019; Dresden. New York: Association for Computing Machinery. https://doi.org/10.1145/3302424.3303982.
47. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. PLoS ONE. 2013. https://doi.org/10.1371/journal.pone.0055811.
48. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574–8. https://doi.org/10.3233/978-1-61499-564-7-574.
49. Beyan O, Choudhury A, van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, et al. Distributed analytics on sensitive medical data: the personal health train. Data Intell. 2020;2:96–107. https://doi.org/10.1162/dint_a_00032.
50. Clinerion Ltd. Patient Network Explorer Solutions [Internet]. Basel: Clinerion; n.d. [Cited 14 June 2021]. Available from https://www.clinerion.com/index/PatientNetworkExplorerSolutions.html.
51. Topaloglu U, Palchuk MB. Using a federated network of real-world data to optimize clinical trials operations. JCO Clin Cancer Inform. 2018;2:1–10. https://doi.org/10.1200/CCI.17.00067.
52. Raisaro JL, Troncoso-Pastoriza JR, Misbach M, Sousa JS, Pradervand S, Missiaglia E, et al. MedCo: enabling secure and privacy-preserving exploration of distributed clinical and genomic data. IEEE/ACM Trans Comput Biol Bioinform. 2018;16:1328–41. https://doi.org/10.1109/TCBB.2018.2854776.
53. Archer DW, Bogdanov D, Lindell Y, Kamm L, Nielsen K, Pagter JI, et al. From keys to databases—real-world applications of secure multi-party computation. Comput J. 2018;61:1749–71. https://doi.org/10.1093/comjnl/bxy090.
54. ISD Services. Use of the National Safe Haven [Internet]. Edinburgh: ISD Services; n.d. [Cited 14 June 2021]. Available from https://www.isdscotland.org/Products-and-Services/EDRIS/Use-of-the-National-Safe-Haven/.
55. ResDAC. CMS Virtual Research Data Center (VRDC) [Internet]. Minneapolis: ResDAC; n.d. [Cited 14 June 2021]. Available from https://www.resdac.org/cms-virtual-research-data-center-vrdc.
56. Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, et al. The shared health research information network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009;16:624–30.
57. Ota S, Cron RQ, Schanberg LE, O'Neil K, Mellins ED, Fuhlbrigge RC, et al. Research priorities in pediatric rheumatology: the childhood arthritis and rheumatology research alliance (CARRA) consensus. Pediatr Rheumatol Online J. 2008;6:5. https://doi.org/10.1186/1546-0096-6-5.
58. Visweswaran S, Becich MJ, D'Itri VS, Sendro ER, MacFadden D, Anderson NR, et al. Accrual to clinical trials (ACT): a clinical and translational science award consortium network. JAMIA Open. 2018;1:147–52. https://doi.org/10.1093/jamiaopen/ooy033.
59. Beenackers MA, Doiron D, Fortier I, Noordzij JM, Reinhard E, Courtin E, et al. MINDMAP: establishing an integrated database infrastructure for research in ageing, mental well-being, and the urban environment. BMC Public Health. 2018;18:158. https://doi.org/10.1186/s12889-018-5031-7.
60. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial

fibrillation. BMC Med Res Methodol. 2020;20:102. https://doi.org/10.1186/s12874-020-00991-3.
61. Almeida J, Trifan A, Hughes N, Rijnbeek P, Oliveira JL. The European health data and evidence network portal [Internet]; Rotterdam: European Health Data & Evidence Network; n.d. [Cited 14 June 2021]. Available from https://www.ohdsi-europe.org/images/symposium-2019/posters/30_Alina_Trifan.pdf.
62. Deist TM, Dankers FJWM, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, et al. Distributed learning on 20 000+ lung cancer patients—the personal health train. Radiother Oncol. 2020;144:189–200. https://doi.org/10.1016/j.radonc.2019.11.019.
63. Turk MA, Landes SD, Formica MK, Goss KD. Intellectual and developmental disability and COVID-19 case-fatality trends: TriNetX analysis. Disabil Health J. 2020;13:100942. https://doi.org/10.1016/j.dhjo.2020.100942.
64. Raisaro JL, Marino F, Troncoso-Pastoriza J, Beau-Lejdstrom R, Bellazzi R, Murphy R, et al. SCOR: a secure international informatics infrastructure to investigate COVID-19. J Am Med Inform Assoc. 2020;11:1721–6. https://doi.org/10.1093/jamia/ocaa172.
65. Laud P, Pankova A. Privacy-preserving record linkage in large databases using secure multiparty computation. BMC Med Genomics. 2018;11:84. https://doi.org/10.1186/s12920-018-0400-8.
66. Platt R, Lieu T. Data enclaves for sharing information derived from clinical and administrative data. JAMA. 2018;320:753–4. https://doi.org/10.1001/jama.2018.9342.
67. Mesa-Eguiagaray I, Wild SH, Rosenberg PS, Bird SM, Brewster DH, et al. Molecular subtypes: a population-based study of Scottish cancer registry data. Br J Cancer. 1997. https://doi.org/10.1038/s41416-020-0938-z.
68. ResDAC. CMS Virtual Research Data Center (VRDC) FAQ [Internet]. Minneapolis: ResDAC; n.d. [Cited 14 June 2021]. Available from https://www.resdac.org/cms-virtual-research-data-center-vrdc-faqs.
69. Baik SH, Kury FSP, McDonald CJ. Risk of Alzheimer's disease among senior medicare beneficiaries treated with androgen deprivation therapy for prostate cancer. J Clin Oncol. 2017;35:3401–9. https://doi.org/10.1200/JCO.2017.72.6109.
70. Milne BJ, Atkinson J, Blakely T, Day H, Douwes J, Gibb S, et al. Data resource profile: The New Zealand integrated data infrastructure (IDI). Int J Epidemiol. 2019;48:677–677e. https://doi.org/10.1093/ije/dyz014.
71. Foster I. Research infrastructure for the safe analysis of sensitive data. Ann Am Acad Pol Soc Sci. 2018;675:102–20. https://doi.org/10.1177/0002716217742610.
72. Alter G, Gonzalez R. Responsible practices for data sharing. Am Psychol. 2018;73:146–56. https://doi.org/10.1037/amp0000258.
73. Arbuckle L, Ritchie F. The five safes of risk-based anonymization. IEEE Secur Privacy. 2019;17:84–9. https://doi.org/10.1109/MSEC.2019.2929282.
74. Knoppers BM. Framework for responsible sharing of genomic and health-related data. HUGO J. 2014. https://doi.org/10.1186/s11568-014-0003-1.
75. Aziz MMA, Sadat MN, Alhadidi D, Wang S, Jiang X, Brown CL, et al. Privacy-preserving techniques of genomic data-a survey. Brief Bioinform. 2019;20:887–95. https://doi.org/10.1093/bib/bbx139.
76. Mittos A, Malin B, Cristofaro ED. Systematizing genome privacy research: a privacy-enhancing technologies perspective. Proc Privacy Enhancing Technol. 2019;2019:87–107. https://doi.org/10.2478/popets-2019-0006.
77. Naveed M, Ayday E, Clayton EW, Fellay J, Gunter CA, Hubaux JP, et al. Privacy in the genomic era. ACM Comput Surv. 2015. https://doi.org/10.1145/2767007.
78. Thapa C, Camtepe S. Precision health data: requirements, challenges and existing techniques for data security and privacy. Comput Biol Med. 2021;129:104130. https://doi.org/10.1016/j.compbiomed.2020.104130.
79. Toh S. Analytic and data sharing options in real-world multidatabase studies of comparative effectiveness and safety of medical products. Clin Pharmacol Ther. 2020;107:834–42. https://doi.org/10.1002/cpt.1754.
80. Ritchie F, Green E. Frameworks, principles and accreditation in modern data management. Bristol Business School Working Papers in Economics. 2020. https://www2.uwe.ac.uk/faculties/BBS/BUS/Research/BCEF/Frameworks.pdf.
81. Wagner I, Eckhoff D. Technical privacy metrics: a systematic survey. ACM Comput Surv. 2018. https://doi.org/10.1145/3168389.
82. Ohno-Machado L, Vinterbo S, Dreiseitl S. Effects of data anonymization by cell suppression on descriptive statistics and predictive modeling performance. J Am Med Inform Assoc. 2002;9:S115–9. https://doi.org/10.1197/jamia.M1241.
83. Sweeney L. k-anonymity: a model for protecting privacy. Int J Unc Fuzz Knowl Based Syst. 2002;10:557–70. https://doi.org/10.1142/S0218488502001648.
84. Weber GM. Federated queries of clinical data repositories: the sum of the parts does not equal the whole. J Am Med Inform Assoc. 2013;20:e155–61. https://doi.org/10.1136/amiajnl-2012-001299.
85. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using bloom filters. BMC Med Inform Decis Mak. 2009;9:41. https://doi.org/10.1186/1472-6947-9-41.
86. Stammler S, Kussel T, Schoppmann P, Stampe F, Tremper G, Katzenbeisser S, et al. Mainzelliste SecureEpiLinker (MainSEL): privacy-preserving record linkage using secure multi-party computation. Bioinform. 2020. https://doi.org/10.1093/bioinformatics/btaa764.
87. Bonofiglio F, Schumacher M, Binder H. Recovery of original individual person data (IPD) inferences from empirical IPD summaries only: applications to distributed computing under disclosure constraints. Stat Med. 2020;39:1183–98. https://doi.org/10.1002/sim.8470.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

Printing copy of the second publication

BMC Bioinformatics

SOFTWARE

Open Access

EasySMPC: a simple but powerful no-code tool for practical secure multiparty computation

Felix Nikolaus Wirth[1*], Tobias Kussel[2], Armin Müller[1], Kay Hamacher[2] and Fabian Prasser[1]

*Correspondence:
felix-nikolaus.wirth@charite.de

[1] Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Medical Informatics Group, Charitéplatz 1, 10117 Berlin, Germany
[2] Computational Biology and Simulation, TU Darmstadt, Darmstadt, Germany

Abstract

Background: Modern biomedical research is data-driven and relies heavily on the re-use and sharing of data. Biomedical data, however, is subject to strict data protection requirements. Due to the complexity of the data required and the scale of data use, obtaining informed consent is often infeasible. Other methods, such as anonymization or federation, in turn have their own limitations. Secure multi-party computation (SMPC) is a cryptographic technology for distributed calculations, which brings formally provable security and privacy guarantees and can be used to implement a wide-range of analytical approaches. As a relatively new technology, SMPC is still rarely used in real-world biomedical data sharing activities due to several barriers, including its technical complexity and lack of usability.

Results: To overcome these barriers, we have developed the tool *EasySMPC,* which is implemented in Java as a cross-platform, stand-alone desktop application provided as open-source software. The tool makes use of the SMPC method Arithmetic Secret Sharing, which allows to securely sum up pre-defined sets of variables among different parties in two rounds of communication (input sharing and output reconstruction) and integrates this method into a graphical user interface. No additional software services need to be set up or configured, as EasySMPC uses the most widespread digital communication channel available: e-mails. No cryptographic keys need to be exchanged between the parties and e-mails are exchanged automatically by the software. To demonstrate the practicability of our solution, we evaluated its performance in a wide range of data sharing scenarios. The results of our evaluation show that our approach is scalable (summing up 10,000 variables between 20 parties takes less than 300 s) and that the number of participants is the essential factor.

Conclusions: We have developed an easy-to-use "no-code solution" for performing secure joint calculations on biomedical data using SMPC protocols, which is suitable for use by scientists without IT expertise and which has no special infrastructure requirements. We believe that innovative approaches to data sharing with SMPC are needed to foster the translation of complex protocols into practice.

Keywords: Secure multi-party computation, SMPC, Secret sharing, GMW protocol, User experience, No-code, Joint calculations

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Introduction

Biomedical research is becoming increasingly data-driven [1]. To create the large datasets needed to answer precise scientific questions, data needs to be re-used for more than the initial purpose of collection and shared between different actors in the healthcare system and the research community [2–7]. As a consequence, "data sharing" is endorsed by various funding agencies (e.g., [8–10]) and increasingly implemented in practice [11, 12]. The term "data sharing" is used in a variety of ways. In this paper, we use it to refer to joint analyses of data stored at different institutions, which does not necessarily require the exchange of individual-level data. In research, data sharing can enable the generation of new knowledge (e.g., [13]) and also lead to higher citation rates [14, 15]. In addition to the increasing promotion of data sharing, there are also major hurdles to its adoption. Here, data protection and data privacy concerns are a central example (e.g., [7]). However, patients and the public have a positive attitude toward data sharing as long as their privacy is being protected [16–18].

Important laws protecting the privacy of patients and probands include the US Health Insurance Portability and Accountability Act (HIPAA) [19] and the EU General Data Protection Regulation (GDPR) [20]. Re-using or sharing data typically requires either (1) obtaining informed consent or (2) anonymizing the data [21]. However, on the one hand, obtaining consent is often infeasible, e.g., when data is analyzed in retrospect [22]. Anonymization, on the other hand, requires making inherent trade-offs between the degree of protection and the quality and hence utility of output data [23], often rendering individual-level data unsuited for answering medical research questions. As a result, a range of alternative approaches have been developed [24]. One example are distributed data sharing networks, in which no individual-level data, but aggregated results, are being shared amongst the partners to perform various types of joint analyses [25–27]. However, also this approach has limitations, for example when very small patient populations, e.g., with rare diseases, are to be studied, whose data cannot be aggregated [28].

Secure multi-party computation (SMPC) is an emerging cryptographic technology [29–31], which can be used to address the shortcomings of federated data networks. On an abstract level, SMPC protocols provide guarantees comparable to those of a trusted third party, with which the participating parties share their data with [32]. This trusted third party performs joint analyses and sends only the results back to the participants. The involved parties do not directly exchange data with each other and hence no information is being disclosed between them. SMPC can provide exactly the same guarantees by following specific cryptographic protocols that exchange encrypted data between the parties—without a trusted third party being involved. SMPC offers provable security guarantees and clearly stated assumptions. Especially for extremely sensitive information, including various types of biomedical data as targeted in this work, those strong guarantees provide a way to perform distributed analyses that otherwise could not be performed due to data protection challenges.

As a relatively new technology, SMPC has only been implemented for practical data sharing in the last few years [33–35] and it has been argued that this is the case in biomedical research as well [36, 37]. While some examples have been described in the literature, e.g., for survival analyses, genome-wide association studies [38–41], genomic

diagnostics, detection of adverse drug events, or infection numbers during the COVID-19-epidemic [42] (see Section "Comparison with Prior Work"), these are mostly research prototypes or specific implementations of SMPC for specific analyses in the context of specific projects. There are several reasons for the slow adoption of SMPC technologies, amongst which are legal barriers, communication barriers, technical barriers and usability challenges (see "Limitations and future work" section).

Challenges and objectives

In the work described in this paper, we addressed two important barriers—technical complexity and usability—to foster the adoption of SMPC technologies for biomedical data sharing:

1. Technical complexity: To enable distributed analyses of data across institutions, external queries against local IT solutions must be allowed and responses must be returned. This requires the installation of local services and an opening up of institutional firewalls. Both needs to be done with great care, which can lead to high efforts and potentially a reluctance to participate in data sharing networks.
2. Usability: SMPC protocols are typically implemented as command-line applications or provided as programming libraries (e.g., for statistical computing environments), thus addressing technical specialists, data scientists or other SMPC researchers. This makes it difficult for scientists involved in biomedical research projects, such as clinicians, to engage in SMPC-based data sharing.

We tackled these challenges by developing *EasySMPC*, which provides a "no-code solution" for securely performing joint calculations on distributed data using an intuitive graphical application. Moreover, no local services need to be installed and no permissive network configuration is necessary, as the application uses e-mails to exchange data between the participants while executing its protocol. To demonstrate the practicability of our solution, we evaluated its performance in a wide range of data sharing scenarios.

Implementation

Secure multi-party computation

SMPC describes a field of cryptographic techniques concerned with joint computations while maintaining confidentiality guarantees regarding the parties' secret inputs. The field emerged in the 1980s with Andrew Yao's publication of the "Garbled Circuits" protocol [43]. Another widely used SMPC method is the GMW-Protocol [44], which describes a way to securely compute a joint (Boolean) function on the secret inputs of n parties. The underlying Boolean circuit uses only logical AND and XOR operations (that is, it states the function in algebraic normal form).

The GMW protocol can easily be extended to not only operate on Boolean circuits with logical values, but also on Arithmetic circuits with values of a finite ring. The idea of the secret sharing scheme is the same in both variants: generate shares (henceforth called "secret shares") by mixing the secret value with randomness so that the combination of all shares results in the reconstructed secret. In the joint arithmetic computation,

additions can be evaluated locally and multiplications are evaluated using interactive sub-protocols, such as the Gilboa-Multiplication for the two-party case [45].

This arithmetic extension of the GMW protocol, referred to as *Arithmetic Secret Sharing*, is the central method implemented in EasySMPC. For further information, we refer interested readers to Additional file 1 of this paper and to the literature (the book by Evans et al. provides a good starting point [46]).

Design of EasySMPC

General approach

The general idea of *EasySMPC* is to provide a user-friendly tool for making SMPC-based data sharing available through an intuitive interface. EasySMPC uses Arithmetic Secret Sharing over the finite ring $\mathbb{Z}(2^{127}-1)$, that is a ring of integers with $2^{127}-1$ elements. This assures, that for all practical values and number of parties the computation will not be restricted by the size of the finite field.[1] As we only employ addition in this version, the protocol can be evaluated with two rounds of communication: first one round of sending/receiving shares for the values that are to be kept secret (e.g., case numbers of a rare disease in a hospital), hence revealing no information, and then a second round of sending/receiving shares for the intermediate results which can then be recombined to obtain the final result. As an inherent property of this family of secure protocols, this can be implemented without exchanging cryptographic keys in the classical sense during set up or prior to a computation, which is an additional factor contributing to the usability of the tool. Finally, we note that the scheme used by EasySMPC is a "full-threshold" protocol, meaning that it is robust against up to $n-1$ corrupted parties, where n is the total number of participating parties, thus, providing a very high degree of protection.

From the user perspective, EasySMPC uses three concepts: (1) *Studies* are the overarching concept composed of participants, variables and protocol states; (2) *Participants* refer to different people or institutions, such as hospitals, who wish to engage in a common computation. Participants are identified by their name and e-mail address. Each study is initiated by exactly one study creator and involves two or more additional participants; (3) *Variables* refer to the data items that are independently summed up in one data sharing process and which are identified by unique names.

Figure 1 provides an overview of the overall process implemented by EasySMPC and the different steps that users need go through when using the tool.

As depicted, the process consists of two rounds of data exchange: In the first round, meta-data and the shares for the participants' secret values are exchanged. For this purpose, the study initiator creates the study, thereby providing a study name, a list of participants and their contact details as well as the list of (named) variables that will be summed up. The initiator also enters their own secret value for each variable, which will remain confidential. The sharable information is then sent to all other participants. Each participant receives their message, initializes the study and enters their own secret value for each variable, which will also remain confidential. Each participant (apart from the initiator) now sends a message to all other participants to

[1] We note that EasySMPC nevertheless supports the summation of decimal numbers by using a fixed-point representation.

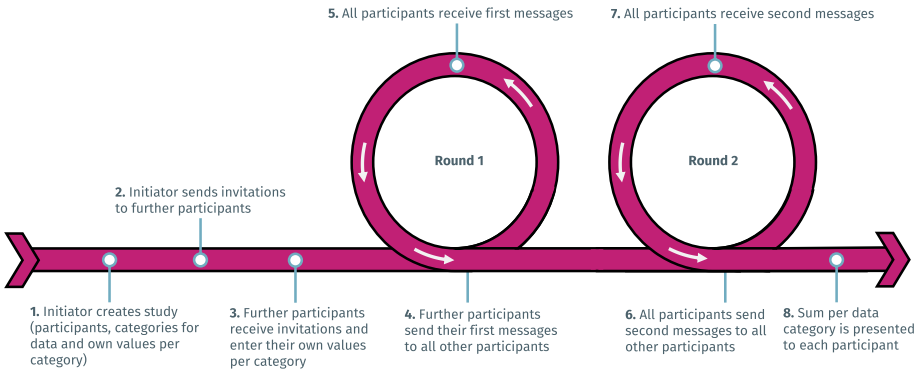


Fig. 1 Overview of the steps in EasySMPC

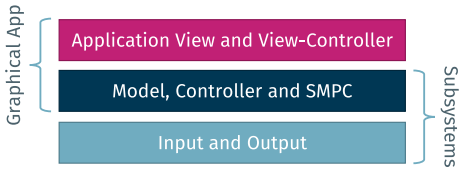


Fig. 2 General architecture of EasySMPC

distribute their respective secret share. Between communication rounds, each party calculates their new secret share locally by summing up the secret shares from round 1. In the second round, the same process is repeated, thereby exchanging the shares of the result. When a participant receives the final message, the result is reconstructed from the secret shares and the resulting sum for each variable across all participants is displayed. With n participants, each user sends and receives $2 \cdot (n-1)$ messages. That is, the number of messages for each participant grows linearly with the number of participants, implying that the *overall* number of messages sent during a calculation grows quadratically.

EasySMPC offers two ways of exchanging messages: (1) in the *semi-manual mode* the users exchange all messages by manually using their preferred e-mail client. The e-mails are, however, pre-generated by EasySMPC and can be imported automatically from the clipboard; (2) in the *automated mode* the participants receive and import the initial message manually. All further messages are exchanged automatically by an e-mail client built into the software.

Architecture and implementation of the software

The architecture of EasySMPC follows the classic model-view-controller approach which is often used to implement applications with graphical user interfaces [47]. An overview of the most important modules is presented in Fig. 2.

EasySMPC is implemented in Java as a cross-platform, stand-alone application that was tested on Windows, MacOS and Linux. The graphical application is built on top of two subsystems, (1) one for cryptographic SMPC operations and (2) one for input- and output as well as data exchange with external applications and the other participants. The application itself consists of a module containing the different user-facing

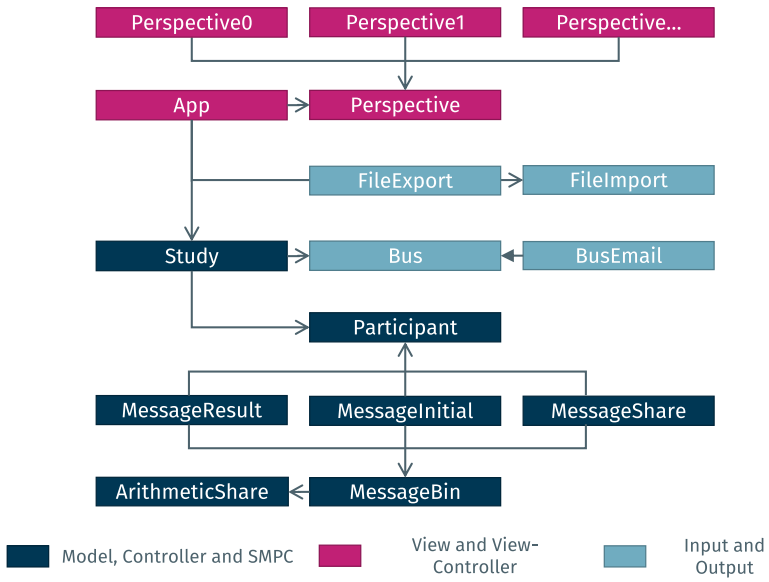


Fig. 3 High-level class diagram

views and perspectives (described in more detail in the following section), as well as parts of the application controller, which is in charge of manipulating the model.

In detail, the three modules are designed as follows: (1) The *Application View and View-Controller* consists of eight different perspectives that reflect the process illustrated in Fig. 1 and guide users through its execution. For the perspectives, highly extendable components based on Java Swing were implemented. (2) The *Model, Controller and SMPC* module *is two-fold:* The module contains (a) the application model holding all data that is needed for executing the protocol and provides methods to safely switch between the states defined in the state machine (see below). Moreover, the module implements (b) the cryptographic Arithmetic Secret Sharing scheme presented in Additional file 1 of this paper. All interactions with this part of the subsystem are performed through the application model. (3) The *Input and Output (I/O) subsystem* provides functionalities for importing data from Excel and CSV files and for sending and receiving data by e-mail. A message can either be sent semi-manually by opening the user's default e-mail client with all relevant fields (recipient, name of study etc.) pre-filled or in a fully automated manner by the I/O subsystem. In both cases the message itself is included in each mail as a Base64 encoded string. Each message contains all relevant metadata including the participants of the calculation, the name of all variables and the current state of the protocol execution, as well as a checksum to detect possible corruptions. Note, that a corrupted message may only lead to an erroneous result but cannot compromise input data privacy. A message can be received semi-manually by copying and pasting data into EasySMPC or be retrieved automatically by the I/O subsystem. In the first case, the application also monitors the user's clipboard and automatically imports all EasySMPC-related messages that are contained in any text copied by the user. In the second case, a bus specifically developed for EasySMPC is used to exchange the data automatically between the different e-mail accounts.

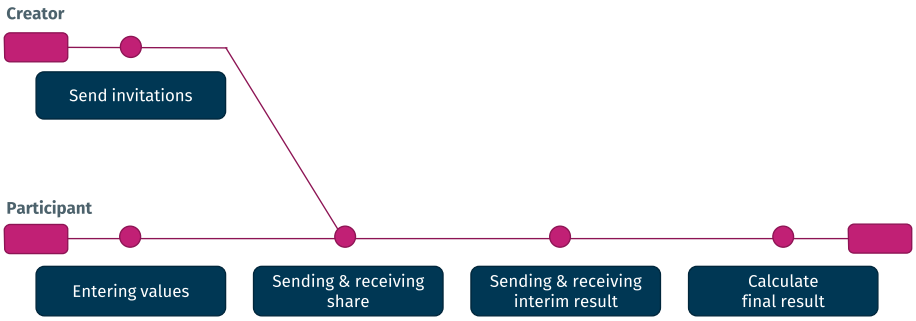


Fig. 4 Application states

For the implementation, Java standard libraries as well as the libraries Jakarta Mail, Apache POI, Commons and Logging were used. Figure 3 displays a high-level class diagram of the software. The class *Study* is central to the execution of calculations through EasySMPC, as it implements the core algorithm. It makes use of further classes in the same module representing *Participants* as well as various types of messages and data used and exchanged. Data exchange is implemented through an abstract *Bus* system of which an implementation using e-mail is included. User interaction is controlled through the *App*, which contains the various perspectives described. It also acts as a mediator between the perspectives, the SMPC algorithm, data exchange and the tool's data import and export capabilities.

As mentioned, a finite state machine makes sure that the cryptographic protocol is followed as needed and that no invalid state transitions are being performed. The states and possible transitions are shown in Fig. 4. The state machine is also the reason why the application model, which handles the current state of the software, also contains parts of the controller. Given the asynchronous nature of data exchange, the API also allows saving the current state of the application at any time, not only after state transitions have been finalized.

Results

Overview of the software

The different perspectives of EasySMPC are shown in Fig. 5. In the example, a common frequency distribution of co-morbidities of patients with Phenylketonuria (PKU), a congenital metabolic disease, is computed with four participating health care institutions. The figure shows the perspectives for (1) initializing a study, (2) sending messages, (3) receiving messages and (4) displaying the result. Similar perspectives that are used for the second round of the protocol have been omitted for brevity.

As can be seen, EasySMPC features a structured and intuitive design, in which data is displayed to the users in tabular form. A progress bar at the top of the application informs the user about the current step in the execution of the protocol. Important actions for the respective step are directly available in each perspective. Further operations, such as loading and saving a project, can be performed via the application menu.

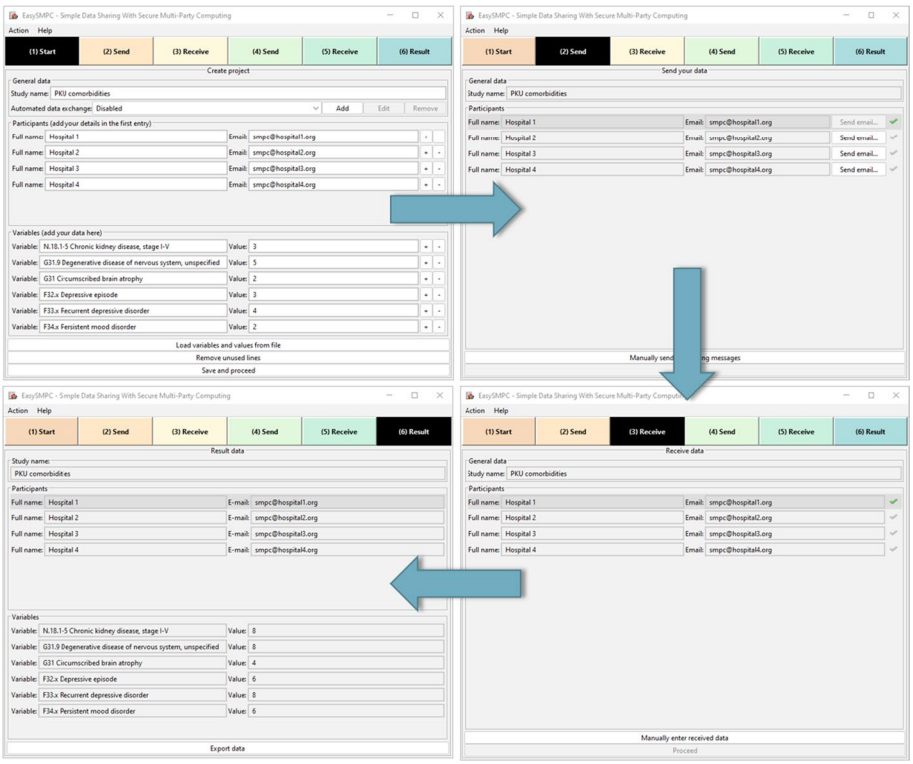


Fig. 5 Perspectives of EasySMPC for (1) initializing a study, (2) sending messages, (3) receiving messages and (4) displaying the result

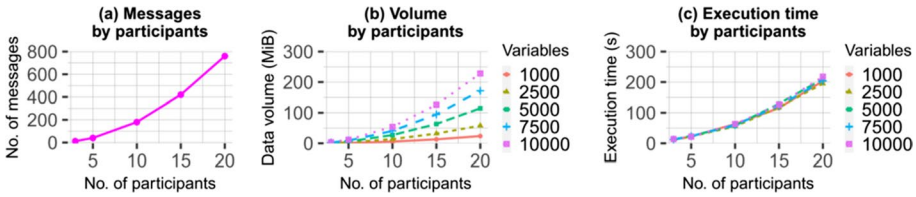


Fig. 6 Experimental results obtained using the default settings

Performance evaluation

To evaluate the performance of EasySMPC we performed a wide range of experiments covering realistic application scenarios. Here we quickly provide an overview of results obtained using the default settings of EasySMPC. For a detailed description of the experimental setup and the results we refer to Additional file 2.

We varied two aspects: (1) the number of participants and (2) number of variables.

Figure 6a shows the total number of messages exchanged when processing the data of a varying number of participants while Fig. 6b, c show the total exchanged data volumes and execution times, which depend on the number of variables summed up as well as the number of participants.

In summary, our experiments confirm that the approach implemented by EasySMPC is feasible even in complex scenarios. The aggregation of 10,000 variables amongst 20 participants can be performed in less than five minutes.

The size of the messages exchanged by EasySMPC depends on the length of the names of the variables and the sizes of its values. The numbers obtained in our experiments show that, in a typical usage scenario, it can be expected that each variable-value-pair can be encoded in approximately 30 bytes (we used 10 random letters for each variable and values in the range of single-precision floating-point numbers). Many mail servers enforce a limit on the maximum size of messages that can be processed. Assuming a conservative limit of 10 Mbyte and based on the data obtained in our experiments this limit would be reached with about 340,000 variables. However, to support scenarios with even more variables, EasySMPC will split up larger messages into several smaller messages. The maximum message size is configurable in the software.

More details on the complexity of the algorithms involved is provided in "Computational complexity" section.

Discussion

Principal results

EasySMPC is a tool that allows summing up values of variables keeping the participants' inputs confidential. To realize this, the software uses an established Arithmetic Secret Sharing protocol.

EasySMPC's innovative aspects lie in the fact that it is very easy to roll out, as no additional effort for installing software services or configuring network interfaces is required and that it offers an intuitive user interface that addresses the needs of non-technical users, such as medical researchers. Through integration into the users' desktop environments and existing e-mail infrastructures, the tool is able to leverage the most common communication channel that is likely to be readily available at sites wanting to engage in a common secure calculation. By using multiple rounds of calculations, several important statistical analyses can be realized (see next section). We have demonstrated its practicability by an extensive evaluation. EasySMPC is released as open-source software under a permissive license and its source code is available online [48].

Supported data analyses

To make EasySMPC as easy to use as possible, the range of supported functionality has been kept to a minimum, focusing on the secure addition of a pre-defined set of variables. However, this basic functionality can be used to perform a range of more complex statistical analyses. For this purpose, different (derived) variables can be processed in multiple cycles, where each cycle is defined as one execution of EasySMPC, i.e., two rounds of sending and receiving messages. An overview of how the most fundamental statistical methods in biomedical research, as identified by Scotch et al. [49], can be implemented with EasySMPC is provided in Table 1.

The table shows that a range of analyses can be performed with one cycle in EasySMPC. Most of these analyses are suited for variables with a nominal level of measurement (indicating that the values have no natural order) and variables with an ordinal scale of measure (indicating that values have a natural order, but no relative distance between values can be expressed). Important examples include the computation of common frequency distributions (already mentioned above) and chi-square tests, where the cells of the relevant contingency table have to be defined a priori and cell counts

Table 1 Example of common statistical methods that can be implemented with EasySMPC

Statistical method	Level of measurement	Input data[a]	Cycles with *EasySMPC*
Frequency distribution	Nominal	Local frequencies per class	1
Chi-square test	Nominal	Local frequencies per cell	1
Quartiles (median, interquartile range)	Ordinal	Local frequencies per class	1
Wilcoxon rank sum test	Ordinal	Local frequencies per class	1
Mean	Interval	Local sum and local count of values	1
Standard deviation (SD)	Interval	Data for mean and local deviation of mean	2
t-test/analysis of variance (ANOVA)[b]	Interval	Local sum, local count of values and local deviation of group mean	2
Correlation coefficient[c]	Interval	SD per variable, co-variance per variable	3

[a] All participants learn the global sum of the data entered locally. No participant learns local values of the other participants

[b] t-test is a special case of the analysis of variance with two groups

[c] Only possible if data for both variables to be correlated are available at the parties (horizontal data distribution)

can be summed up with EasySMPC to derive the final chi-square statistics. For ordinal data, quartiles can be derived from the common frequency distribution. Moreover, an inferential test of two independent distributions, the Wilcoxon rank sum test, can be performed using two common distributions computed with EasySMPC. For variables with an interval scale (indicating a natural order and a relative distance between values), further analyses are supported. For example, a common mean can be calculated by having each participant share a sum of a variable and the number of values, which can be divided with each other after computing common sums. Implementing further statistical analyses will require more than one cycle. For example, the standard deviation of a common distribution can be computed by calculating the mean in a first cycle. In a second cycle, each participant can calculate the variation of its data compared to the global mean. By using the variance computed in the second cycle and the total number of values calculated in the first cycle, the participants can further calculate the total standard deviation. In a third cycle, the total covariance can be computed to investigate a correlation for horizontally distributed data. Analogously, a t-test or analysis of variance can be performed by calculating the mean per group in a first cycle and the variance of local data in relationship to the global mean in a second cycle. When all those common sums are computed, the t-test and analysis of variance (ANOVA) statistics can be calculated.

We note that when an analysis is performed using more than one cycle, more data will be disclosed than when the complete process would have been performed using a tailored SMPC protocol. However, we would like to point out that, as already mentioned above, only aggregated and likely less sensitive data (cf. GDPR Recital 162 (5) [20]) is disclosed in the intermediate results. However, this needs to be carefully analyzed on a case-by-case basis before performing more complex analyses.

Computational complexity

With its actual runtime being highly dependent on the employed (networking) hardware, the asymptotic complexities regarding runtime and space usage are important for evaluating the protocol. EasySMPC employs a SMPC protocol with a constant number of communication rounds and outside of those interactions only non-interactive,

computationally inexpensive additions. This means that EasySMPC's asymptotic runtime complexity is linear in the number of network interactions. The number of messages sent by each participant in a computation with n participants is $2 \cdot (n-1)$ (see also "Design of EasySMPC" section). This also means that it is unlikely that limits of typical mail servers regarding the number of messages that can be sent within a certain timeframe will be reached in calculations with a reasonable number of participants. The *overall* number of messages, which determines runtime performance, is $\mathcal{O}(n^2)$, which is executed in a parallel manner over n concurrent processes (one executed by each participant).

Space complexity, again, is dependent on the number of messages. The messages contain the variable names and values, as well as a small overhead. Each individual message scales linearly in the number of variables. The overall space complexity of EasySMPC therefore is $\mathcal{O}(v \cdot n^2)$ with v being the number of variables, where each participant needs memory of $\mathcal{O}(v \cdot n)$.

Lastly, the consecutive execution of EasySMPC to create the more complex analyses listed in Table 1 (see "Supported data analyses" section) compose linearly, as all examples use the same number of participants and variables for each iteration. As the number of iterations is small in every given case, the incurred small factor can be omitted in an asymptotic complexity analysis.

Comparison with prior work

A number of SMPC protocols and solutions have already been described in the literature that can be used in different areas of biomedical research. For example, Stammler et al. [41] and other authors [50–52] have investigated general secure record-linkage processes [53]. Moreover, El Emam et al. describe a protocol for the secure linkage of data for surveillance registries [54]. Several works describe the application of SMPC techniques for specific use cases in biomedical research. Examples include methods for conducting drug-target interaction assessments [55, 56], drug screening [57], genome-wide association studies [38, 39, 58–63] and genomic diagnostics [64]. Other works propose the application of SMPC techniques to realize specific statistical methods allowing biomedical data analyses, such as (1) the calculation of Kaplan–Meier estimators [65, 66], (2) linear [67] or (3) logistic [68–71] regression analyses and k-means clustering [72]. In addition, there are generic frameworks that can be used as a basis for implementing specific SMPC algorithms. Important examples include technical programming libraries and environments such as Sharemind MPC [73], FRESCO [74], ABY [75], MOTION [76] or MP-SPDZ [77] and generic data sharing infrastructures, such as MedCo [78] or FAMHE [79]. Tools that specifically target usability are also a hot topic in the biomedical field (see, e.g., [80, 81] for recent examples).

The papers cited in the first three areas describe complex algorithms which have been developed for a particular purpose. EasySMPC, on the other hand, follows a different strategy and supports a generic functionality optimized for usability by people that are not IT specialists. Moreover, we note that EasySMPC is not a research prototype but has been designed for real-world applications. The same is true for MedCo and FAMHE, which provide more comprehensive functionalities than EasySMPC.

However, the efforts required to install, configure and maintain these solutions is relatively high, while EasySMPC was designed to be as easy as possible to install and use.

Limitations and future work

The current restriction of EasySMPC to addition and subtraction is a major limitation of the software. While, as we have shown, this basic functionality can be used to implement a range of analyses, this can be cumbersome, as several independent rounds need to be performed. In future versions of the tool, we plan to add support for additional basic operations as well as more complex data analyses. On the user interface level, we plan to maintain EasySMPC's usability by using a spreadsheet-like approach for entering data and displaying results.

In addition to the controlled experiments presented in this paper, we have also performed feasibility evaluations with EasySMPC in a real-world setting involving several hospitals from the German CORD project for research on rare diseases. While EasySMPC worked very well in all of those settings, the use of e-mail as a communication infrastructure resulted in some limitations. One example is that common mail servers may flag communication as spam if a very large number of messages is exchanged due to a large number of participants being involved. To also support such use cases, work is currently underway to extend the bus functionality of EasySMPC to other common communication technologies.

On the security and privacy-side, some trade-offs had to be made. First, the different parties are only authenticated via access to the e-mail accounts, meaning that a man in the middle attack could be performed and the integrity of the calculation cannot be guaranteed. However, this does not affect the confidentiality of the data entered by the participants, since the employed protocol is proven to be secure [44]. Thus, in the worst case, an attacker might maliciously change the calculated results, but is never able to obtain the input data of other participants. Moreover, like many other SMPC solutions [34], EasySMPC provides a safe setting for processing data but does not necessarily guarantee that the output data is also protected (see also "Supported data analyses" section). In future work, we plan to address these issues by integrating more comprehensive authentication mechanisms and methods for providing safe outputs, such as Differential Privacy [82].

Finally, there are a few general barriers to the further adoption of SMPC methods that are not specific to EasySMPC. For example, Tõldsepp et al. [83] identified the following important challenges that also apply to our software: (1) legal frameworks often do not consider SMPC, methods which in turn leads to legal uncertainties (see also [37]), (2) it can be challenging to explain and communicate the properties of SMPC to relevant stakeholders (e.g., Institutional Review Boards (IRBs) or ethics committees; see also [37, 46, 84]), (3) users may misuse SMPC technologies leading to additional risks in the *honest but curious* attacker model typically assumed (see also [85]) and (4) data analysts might find it difficult to analyze data they cannot access directly (see also [46, 86]). By developing EasySMPC which makes such technologies available to a broader audience and more use cases, we hope to be able to contribute to overcoming these barriers.

Conclusions

In this paper we have presented EasySMPC, a user-friendly graphical application supporting the secure analysis of distributed data across multiple institutions without requiring IT expertise. Although SMPC methods are considered a break-through technology for data-driven medical research, they are not in widespread use to date and implementing them can be associated with major hurdles. We believe that innovative no-code approaches to secure data sharing, as the one presented in this paper, can foster the translation of more complex protocols into practice.

Availability and requirements

Project name: EasySMPC. Project home page: https://github.com/prasser/easy-smpc. Operating system(s): Platform independent. Programming language: Java. Other requirements: Java 14 or higher. License: Apache 2.0. Any restrictions to use by non-academics: none.

Abbreviations

ANOVA	Analysis of variance
GDPR	General data protection regulation
HIPAA	Health Insurance Portability and Accountability Act
I/O	Input and output
IRB	Institutional Review Board
OT	Oblivious transfer
PKU	Phenylketonuria
SD	Standard deviation
SMPC	Secure multi-party computation
XOR	Exclusively-OR

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-05044-8.

Additional file 1. Microsoft Word format describes the employed SMPC method in detail.

Additional file 2. Microsoft Word format contains the detailed results of the performance evaluation.

Acknowledgements

We thank our anonymous reviewers for the constructive feedback

Author contributions

TK designed and developed the cryptographic part of the software. FP designed the architecture of the non-cryptographic part of the software. FNW, AM and FP implemented the Graphical User Interface, FNW and FP developed and evaluated the bus functionality. FNW, TK, FP, AM and KH drafted the manuscript. FP and KH revised the manuscript. All authors have read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—SFB 1119-236615297 and by the German Ministry of Education and Research through the project CORD-MI (funding #01ZZ1911F). The funders had no role in the design of the study, data collection and analysis, writing of the manuscript, or the decision to publish.

Availability of data and materials

The performance evaluation dataset generated and analyzed during the current study is available in the GitHub repository of the performance evaluation, https://github.com/fnwirth/easy-smpc-performance-evaluation.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 31 May 2022 Accepted: 8 November 2022
Published online: 09 December 2022

References

1. Munevar S. Unlocking Big Data for better health. Nat Biotechnol. 2017;35:684–6. https://doi.org/10.1038/nbt.3918.
2. Gewin V. Data sharing: an open mind on open data. Nature. 2016;529:117–9. https://doi.org/10.1038/nj7584-117a.
3. Merson L, Gaye O, Guerin PJ. Avoiding data dumpsters-toward equitable and useful data sharing. N Engl J Med. 2016;374:2414–5. https://doi.org/10.1056/NEJMp1605148.
4. Taichman DB, Backus J, Baethge C, Bauchner H, de Leeuw PW, Drazen JM, et al. Sharing clinical trial data: a proposal from the International Committee of Medical Journal Editors. N Engl J Med. 2016;374:384–6. https://doi.org/10.1056/NEJMe1515172.
5. Carr D, Littler K. Sharing research data to improve public health. J Empir Res Hum Res Ethics. 2015;10:314–6. https://doi.org/10.1177/1556264615593485.
6. Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. Nat Biotechnol. 2018;36:391–2. https://doi.org/10.1038/nbt.4128.
7. Villanueva AG, Cook-Deegan R, Koenig BA, Deverka PA, Versalovic E, McGuire AL, et al. Characterizing the biomedical data-sharing landscape. J Law Med Ethics. 2019;47:21–30. https://doi.org/10.1177/1073110519840481.
8. Pilat D, Fukasaku Y. OECD principles and guidelines for access to research data from public funding. Data Sci J. 2007;6:OD4–11. https://doi.org/10.2481/dsj.6.OD4.
9. Walport M, Brest P. Sharing research data to improve public health. Lancet. 2011;377:537–9. https://doi.org/10.1016/S0140-6736(10)62234-9.
10. Australien Government—National Health and Medical Research Council. Open Access Policy 2018. https://www.nhmrc.gov.au/file/15242/download?token=rgNjnh0B. Accessed 29 July 2022.
11. Institute of Medicine (US). Sharing Clinical Research Data: Workshop Summary. Washington: The National Academies Press; 2013.
12. Hulsen T. Sharing is caring-data sharing initiatives in healthcare. Int J Environ Res Public Health. 2020. https://doi.org/10.3390/ijerph17093046.
13. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet. 2019;51:237–44. https://doi.org/10.1038/s41588-018-0307-5.
14. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. PLoS ONE. 2007;2:e308. https://doi.org/10.1371/journal.pone.0000308.
15. Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. PeerJ. 2013;1:e175. https://doi.org/10.7717/peerj.175.
16. Kim KK, Joseph JG, Ohno-Machado L. Comparison of consumers' views on electronic data sharing for healthcare and research. J Am Med Inform Assoc. 2015;22:821–30. https://doi.org/10.1093/jamia/ocv014.
17. Aitken M, de St JJ, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics. 2016;17:73. https://doi.org/10.1186/s12910-016-0153-x.
18. Kalkman S, van Delden J, Banerjee A, Tyl B, Mostert M, van Thiel G. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. J Med Ethics. 2019. https://doi.org/10.1136/medethics-2019-105651.
19. United States Congress. Health insurance portability and accountability act of 1996. Public Law. 1996;104:191.
20. Regulation GDP. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. Off J Eur Union (OJ). 2016;59:294.
21. Emam KE, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ. 2015. https://doi.org/10.1136/bmj.h1139.
22. Williams G, Pigeot I. Consent and confidentiality in the light of recent demands for data sharing. BIOM J. 2017;59:240–50. https://doi.org/10.1002/bimj.201500044.
23. Prasser F, Eicher J, Spengler H, et al. Flexible data anonymization using ARX—current status and challenges ahead. Softw Pract Exp. 2020;50:1277–304. https://doi.org/10.1002/spe.2812.
24. Wirth FN, Meurers T, Johns M, Prasser F. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. BMC Med Inform Decis Mak. 2021;21:242. https://doi.org/10.1186/s12911-021-01602-x.
25. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574–8. https://doi.org/10.3233/978-1-61499-564-7-574.
26. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. J Am Med Inform Assoc. 2014;21:576–7. https://doi.org/10.1136/amiajnl-2014-002864.
27. Topaloglu U, Palchuk MB. Using a federated network of real-world data to optimize clinical trials operations. JCO Clin Cancer Inform. 2018;2:1–10. https://doi.org/10.1200/CCI.17.00067.
28. MacLeod H, Abbott J, Patil S. Small data privacy protection: an exploration of the utility of anonymized data of people with rare diseases. In: Mark G, Fussell S, editors. WISH'17. Proceedings of the 2017 workshop on interactive

systems in healthcare. May 6–11, 2017; Colorado. Washington: Association for Computing Machinery; 2017, p. 3059–64. https://doi.org/10.1145/3027063.3108900.
29. Berger B, Cho H. Emerging technologies towards enhancing privacy in genomic data sharing. Genome Biol. 2019;20:128. https://doi.org/10.1186/s13059-019-1741-0.
30. Telenti A, Jiang X. Treating medical data as a durable asset. Nat Genet. 2020;52:1005–10. https://doi.org/10.1038/s41588-020-0698-y.
31. Gartner Research. Hype Cycle for Privacy 2020. 2020. https://www.gartner.com/en/documents/3987903/hype-cycle-for-privacy-2020. Accessed 29 July 2022.
32. Canetti R. Security and composition of multiparty cryptographic protocols. J Cryptology. 2000;13:143–202. https://doi.org/10.1007/s001459910006.
33. Choi JI, Butler KRB. Secure multiparty computation and trusted hardware: examining adoption challenges and opportunities. Secur Commun Netw. 2019. https://doi.org/10.1155/2019/1368905.
34. Lindell Y. Secure multiparty computation. Commun ACM. 2021;64:86–96. https://doi.org/10.1145/3387108.
35. Hastings M, Hemenway B, Noble D, Zdancewic S. Sok: general purpose compilers for secure multi-party computation. In: Gondree M, editor. 2019 IEEE symposium on security and privacy (SP); 20–22 May 2019; San Francisco. New York: IEEE; 2019, p. 1220–37. https://doi.org/10.1109/SP.2019.00028.
36. Dankar FK, Madathil N, Dankar SK, Boughorbel S. Privacy-preserving analysis of distributed biomedical data: designing efficient and secure multiparty computations using distributed statistical learning theory. JMIR Med Inform. 2019;7:e12702. https://doi.org/10.2196/12702.
37. Veeningen M, Chatterjea S, Horváth AZ, Spindler G, Boersma E, van der Spek P, et al. Enabling analytics on sensitive medical data with secure multi-party computation. Stud Health Technol Inform. 2018;247:76–80.
38. Tkachenko O, Weinert C, Schneider T, Hamacher K. Large-scale privacy-preserving statistical computations for distributed genome-wide association studies. In: Kim J, Ahn G-J, Kim S, editors. ASIACCS '18: Proceedings of the 2018 on Asia conference on computer and communications security; 4 June 2018; Incheon. Washington: Association for Computing Machinery; 2018, p. 221–35.
39. Demmler D, Hamacher K, Schneider T, Stammler S. Privacy-preserving whole-genome variant queries. In: Capkun S, Chow SSM, editors. CANS 2017: cryptology and network security—16th international conference; 29 November–2 December 2017. Berlin: Springer; 2017. p. 71–92.
40. Karvelas N, Peter A, Katzenbeisser S, Tews E, Hamacher K. Privacy-preserving whole genome sequence processing through proxy-aided ORAM. In: Ahn G-J, Datta A, editors. WPES '14: Proceedings of the 13th workshop on privacy in the Electronic Society; 3 November 2014; Scottsdale. New York: Association for Computing Machinery; 2014, p. 1–10.
41. Stammler S, Kussel T, Schoppmann P, Stampe F, Tremper G, Katzenbeisser S, et al. Mainzelliste SecureEpiLinker (MainSEL): privacy-preserving record linkage using secure multi-party computation. Bioinformatics. 2022;38:1657–68. https://doi.org/10.1093/bioinformatics/btaa764.
42. Hamacher K, Kussel T, von Landesberger T, Baumgartl T, Höhn M, Scheithauer S, et al. Fallzahlen Re-Identifikation und der technische Datenschutz. DuD. 2022;46:143–8. https://doi.org/10.1007/s11623-022-1579-6.
43. Yao AC-C. How to generate and exchange secrets. SFCS '86: proceedings of the 27th annual symposium on foundations of computer science; 27–29 October 1986. Washington: IEEE Computer Society; 1986, p. 162–7. https://doi.org/10.1109/SFCS.1986.
44. Micali S, Goldreich O, Wigderson A. How to play any mental game. In: Aho A, editor. STOC '87: Proceedings of the nineteenth ACM symposium on theory of computing; 25–27 May 1987; New York: Association for Computing Machinery; 1987, p. 218–29. https://doi.org/10.1145/28395.28420.
45. Gilboa N. Two party RSA key generation. In: Wiener M, editor. CRYPTO 99: 19th annual international cryptology conference; 15–19 August 1999; Santa Barbara. Berlin, Heidelberg: Springer; 1999, p. 116–29. https://doi.org/10.1007/3-540-48405-1_8.
46. Evans D, Kolesnikov V, Rosulek M. A pragmatic introduction to secure multi-party computation. Foundations and trends; 2018. https://doi.org/10.1561/3300000019.
47. Krasner GE, Pope ST. A description of the model-view-controller user interface paradigm in the smalltalk-80 system. J Op Prog. 1988;1:26–49.
48. Wirth FN, Kussel T, Müller A, Hamacher K, Prasser F. EasySMPC implementation 2022. https://github.com/prasser/easy-smpc. Accessed 29 July 2022.
49. Scotch M, Duggal M, Brandt C, Lin Z, Shiffman R. Use of statistical analysis in the biomedical informatics literature. J Am Med Inform Assoc. 2010;17:3–5. https://doi.org/10.1197/jamia.M2853.
50. Chen F, Jiang X, Wang S, Schilling LM, Meeker D, Ong T, et al. Perfectly secure and efficient two-party electronic-health-record linkage. IEEE Internet Comput. 2018;22:32–41. https://doi.org/10.1109/MIC.2018.112102542.
51. Lazrig I, Ong TC, Ray I, Ray I, Jiang X, Vaidya J. Privacy preserving probabilistic record linkage without trusted third party. In: McCanny, John, editor. PST2018: Proceedings of the 16th annual conference on privacy, security and trust; 28 - 30 August 2018; Belfast. Washington: IEEE Computer Society; 2018, p. 1–10. https://doi.org/10.1109/PST.2018.8514192.
52. Laud P, Pankova A. Privacy-preserving record linkage in large databases using secure multiparty computation. BMC Med Genomics. 2018;11:84. https://doi.org/10.1186/s12920-018-0400-8.
53. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc. 1969;64:1183–210. https://doi.org/10.1080/01621459.1969.10501049.
54. El Emam K, Samet S, Hu J, Peyton L, Earle C, Jayaraman GC, et al. A protocol for the secure linking of registries for HPV surveillance. PLoS ONE. 2012;7:e39915. https://doi.org/10.1371/journal.pone.0039915.
55. Hie B, Cho H, Berger B. Realizing private and practical pharmacological collaboration. Science. 2018;362:347–50. https://doi.org/10.1126/science.aat4807.
56. Ma R, Li Y, Li C, Wan F, Hu H, Xu W, et al. Secure multiparty computation for privacy-preserving drug discovery. Bioinformatics. 2020;36:2872–80. https://doi.org/10.1093/bioinformatics/btaa038.

57. Shimizu K, Nuida K, Arai H, Mitsunari S, Attrapadung N, Hamada M, et al. Privacy-preserving search for chemical compound databases. BMC Bioinform. 2015;16:S6. https://doi.org/10.1186/1471-2105-16-S18-S6.
58. Bonte C, Makri E, Ardeshirdavani A, Simm J, Moreau Y, Vercauteren F. Towards practical privacy-preserving genome-wide association study. BMC Bioinform. 2018;19:537. https://doi.org/10.1186/s12859-018-2541-3.
59. Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty computation. Nat Biotechnol. 2018;36:547–51. https://doi.org/10.1038/nbt.4108.
60. Lu W-J, Yamada Y, Sakuma J. Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption. BMC Med Inform Decis Mak. 2015;15(Suppl 5):S1. https://doi.org/10.1186/1472-6947-15-S5-S1.
61. Kuo T-T, Jiang X, Tang H, Wang X, Bath T, Bu D, et al. iDASH secure genome analysis competition 2018: blockchain genomic data access logging, homomorphic encryption on GWAS, and DNA segment searching. BMC Med Genomics. 2020;13:98. https://doi.org/10.1186/s12920-020-0715-0.
62. Kamm L, Bogdanov D, Laur S, Vilo J. A new way to protect privacy in large-scale genome-wide association studies. Bioinformatics. 2013;29:886–93. https://doi.org/10.1093/bioinformatics/btt066.
63. Franz M, Deiseroth B, Hamacher K, Jha S, Katzenbeisser S, Schröder H. Towards secure bioinformatics services. In: Danezis G, editor. FC 2011: financial cryptography and data security—15th international conference; March 4 2011; Gros Islet. Berlin: Springer; 2011, p. 276–83. https://doi.org/10.1007/978-3-642-27576-0.
64. Jagadeesh KA, Wu DJ, Birgmeier JA, Boneh D, Bejerano G. Deriving genomic diagnoses without revealing patient genomes. Science. 2017;357:692–5. https://doi.org/10.1126/science.aam9710.
65. Vogelsang L, Lehne M, Schoppmann P, Prasser F, Thun S, Scheuermann B, et al. A secure multi-party computation protocol for time-to-event analyses. Stud Health Technol Inform. 2020;270:8–12. https://doi.org/10.3233/SHTI200112.
66. von Maltitz M, Ballhausen H, Kaul D, Fleischmann DF, Niyazi M, Belka C, et al. A privacy-preserving log-rank test for the kaplan-meier estimator with secure multiparty computation: algorithm development and validation. JMIR Med Inform. 2021;9:e22158. https://doi.org/10.2196/22158.
67. Sadat MN, Jiang X, Aziz MMA, Wang S, Mohammed N. Secure and efficient regression analysis using a hybrid cryptographic framework: development and evaluation. JMIR Med Inform. 2018;6:e14. https://doi.org/10.2196/medinform.8286.
68. El Emam K, Samet S, Arbuckle L, Tamblyn R, Earle C, Kantarcioglu M. A secure distributed logistic regression protocol for the detection of rare adverse drug events. J Am Med Inform Assoc. 2013;20:453–61. https://doi.org/10.1136/amiajnl-2011-000735.
69. Lu Y, Zhou T, Tian Y, Zhu S, Li J. Web-based privacy-preserving multicenter medical data analysis tools via threshold homomorphic encryption: design and development study. J Med Internet Res. 2020;22:e22555. https://doi.org/10.2196/22555.
70. Shi H, Jiang C, Dai W, Jiang X, Tang Y, Ohno-Machado L, et al. Secure multi-pArty computation grid logistic regression (SMAC-GLORE). BMC Med Inform Decis Mak. 2016;16:89. https://doi.org/10.1186/s12911-016-0316-1.
71. De Cock M, Dowsley R, Nascimento ACA, Railsback D, Shen J, Todoki A. High performance logistic regression for privacy-preserving genome analysis. BMC Med Genomics. 2021;14:23. https://doi.org/10.1186/s12920-020-00869-9.
72. Spini G, van Heesch M, Veugen T, Chatterjea S. Private hospital workflow optimization via secure k-means clustering. J Med Syst. 2020;44:8. https://doi.org/10.1007/s10916-019-1473-4.
73. Archer DW, Bogdanov D, Lindell Y, Kamm L, Nielsen K, Pagter JI, et al. From keys to databases—real-world applications of secure multi-party computation. Comput J. 2018;61:1749–71. https://doi.org/10.1093/comjnl/bxy090.
74. Alexandra Institute. FRESCO—a framework for efficient secure computation 2021. https://github.com/aicis/fresco. Accessed 29 July 2022.
75. Demmler D, Schneider T, Zohner M. ABY-A framework for efficient mixed-protocol secure two-party computation. NDSS '15: network and distributed system security symposium; 8–11 February 2015; San Diego. San Diego: NDSS; 2015. https://doi.org/10.14722/ndss.2015.23113.
76. Braun L, Demmler D, Schneider T, Tkachenko O. MOTION—a framework for mixed-protocol multi-party computation. IACR Cryptol EPrint Arch 2020. p.1137. https://doi.org/10.1145/3490390.
77. Keller M. MP-SPDZ: A versatile framework for multi-party computation. In: Ligatti J, Ou X, editors. Proceedings of the 2020 ACM SIGSAC conference on computer and communications security; 9–13 November 2020; virtual. New York: Association for Computing Machinery; 2020, p. 1575–90. https://doi.org/10.1145/3372297.3417872.
78. Raisaro JL, Troncoso-Pastoriza JR, Misbach M, Sousa JS, Pradervand S, Missiaglia E, et al. MedCo: enabling secure and privacy-preserving exploration of distributed clinical and genomic data. IEEE/ACM Trans Comput Biol Bioinform. 2018;16:1328–41. https://doi.org/10.1109/TCBB.2018.2854776.
79. Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, Cuendet MA, Sousa JS, Cho H, et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Nat Commun. 2021;12:5910. https://doi.org/10.1038/s41467-021-25972-y.
80. Zhou Y, Leung S-W, Mizutani S, Takagi T, Tian Y-S. MEPHAS: an interactive graphical user interface for medical and pharmaceutical statistical analysis with R and Shiny. BMC Bioinform. 2020;21:183. https://doi.org/10.1186/s12859-020-3494-x.
81. Koile D, Cordoba M, de Sousa SM, Kauffman MA, Yankilevich P. GenIO: a phenotype-genotype analysis web server for clinical genomics of rare diseases. BMC Bioinform. 2018;19:25. https://doi.org/10.1186/s12859-018-2027-3.
82. Dwork C. Differential privacy: a survey of results. In: Agrawal M, Du D, Duan Z, Li A, editors. TAMC 2008: theory and applications of models of computation 5th international conference; 25–29 April 2008; Xi'an. Berlin: Springer; 2008, p. 1–19. https://doi.org/10.1007/978-3-540-79228-4_1.
83. Tõldsepp K, Pruulmann-Vengerfeldt P, Laud P. Usable and efficient secure multiparty computation—requirements specification based on the interviews. Deliverables in usable and efficient secure multiparty computation UaESMC) Research Project 2015. http://uaesmc.cyber.ee/files/d12final.pdf. Accessed 29 July 2022.
84. Bogdanov D, Kamm L, Laur S, Pruulmann-Vengerfeldt P. Secure multi-party data analysis: end user validation and practical experiments. IACR Cryptol EPrint Arch. 2013. https://eprint.iacr.org/2013/826.pdf. Accessed 29 July 2022.

85. Paverd AJ, Martin A, Brown I. Modelling and automatically analysing privacy properties for honest-but-curious adversaries. University of Oxford 2014. https://www.cs.ox.ac.uk/people/andrew.paverd/casper/casper-privacy-report.pdf. Accessed 29 July 2022.
86. Desai T, Ritchie F, Welpton R. Five safes: designing data access for research. 2016. https://doi.org/10.13140/RG.2.1.3661.1604.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

Curriculum vitae

For data protection reasons, my CV will not be published in the electronic version of my work.

Publication list

1) Johns M, Meurers T, **Wirth FN**, Haber AC, Müller A, Halilovic M, Balzer F, Prasser F. Data Provenance in Biomedical Research: Scoping Review. J Med Internet Res. 2023 Mar 27;25:e42289. doi: 10.2196/42289. PMID: 36972116.
   Impact factor: 7.076
2) **Wirth FN**, Kussel T, Müller A, Hamacher K, Prasser F. EasySMPC: a simple but powerful no-code tool for practical secure multiparty computation. BMC Bioinformatics. 2022 Dec 9;23(1):531. doi: 10.1186/s12859-022-05044-8. PMID: 36494612.
   Impact factor: 3.169
3) **Wirth FN**, Meurers T, Johns M, Prasser F. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. BMC Med Inform Decis Mak. 2021 Aug 12;21(1):242. doi: 10.1186/s12911-021-01602-x. PMID: 34384406.
   Impact factor: 2.317
4) Johns M, Müller A, **Wirth FN**, Prasser F. A Comprehensive Portal for Clinical and Translational Data Warehouses. Stud Health Technol Inform. 2021 May 27;281:462-466. doi: 10.3233/SHTI210201. PMID: 34042786.
5) **Wirth FN**, Johns M, Meurers T, Prasser F. Citizen-Centered Mobile Health Apps Collecting Individual-Level Spatial Data for Infectious Disease Management: Scoping Review. JMIR Mhealth Uhealth. 2020 Nov 10;8(11):e22594. doi: 10.2196/22594. PMID: 33074833.
   Impact factor: 4.73

Acknowledgments

An erster Stelle möchte ich meinen Dank an meine Betreuer Prof. Fabian Prasser und Prof. Dominik Seelow für ihre Unterstützung richten. Ohne das kontinuierliche Feedback und die Unterstützung bei der Betreuung wäre diese Arbeit nicht möglich gewesen.

Ebenfalls sehr dankbar bin ich für die Mitwirkung meiner Ko-Autoren Marco Johns, Thierry Meurers, Armin Müller und Tobias Kussel.

Für jederzeit vorhandene Unterstützung – insbesondere während der Promotion und auch sonst - danke ich zu aller erst meinen Eltern Nicola und Thomas sowie meiner Tante Claudia, meinem Onkel Martin, meinem langjährigen Unterstützer Jörg und in liebevoller Erinnerung meiner Großmutter Hannelore. Ebenfalls sehr dankbar bin ich für meine „Wahlverwandtschaften“ mit Kili, Felix, Rike, Alex, Sophie, Lisa und Hanna. Danke euch allen für alles zu aller Zeit!