ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Why Johnny Can't Compute Securely: Exploring the Gap between Threat Models and Stakeholder Concerns

Semester Project

Lena Csomor

September 18, 2021

Advisors: Prof. Dr. Kenny Paterson, Alexander Viand

Applied Cryptography Group
Institute of Information Security
Department of Computer Science, ETH Zürich

Abstract

Secure computation techniques like Secure Multi-Party Computation, Fully Homomorphic Encryption, and Zero-Knowledge Proofs have made significant progress in performance and feature variety in recent years. Still, we cannot observe a substantial rise in its use in real-life applications. In this thesis, we investigate possible reasons for the slow adoption of secure computation by exploring the gap between the existing cryptographic state-of-the-art and the requirements of real-world stakeholders.

We focus our attention on applications of secure computation in (Swiss) health care and investigate *(i)* the current state of data collection, sharing, and processing in the domain of health care and medical research, *(ii)* the extent to which existing secure computation approaches in research and industry solve the needs of the medical community, and finally *(iii)* what would be required to close the gaps between the current capabilities offered by secure computation on one hand and the requirements of medical applications on the other hand.

Our exploration reveals several widespread challenges and issues stemming from standard medical data sharing practices, hindering and restricting technological advances. We explore existing secure computation approaches in research and industry to see how they could mitigate these restrictions to enable new approaches to, e.g., longitudinal medical study. Our analysis highlights promising prospects but also shows that current secure computation approaches alone are not sufficient to resolve the issue. Finally, we propose possible solutions to close this gap, showing that rather than requiring novel cryptographic primitives, a more pragmatic attitude and a shift to client-centered product development in secure computation providers seem crucial for future advancements.

Acknowledgments

I thank my supervisor Alexander Viand for his continued support, our constructive and fun meetings, and the freedom he gave me while writing this thesis. This enabled me to pursue topics I am very passionate about and to get to the root of exciting problems.

Thanks to Prof. Kenny Paterson, I was allowed to discover the fascinating field of secure computation while exploring the intersection between medicine and technology.

A special thank you goes to my boyfriend Noah Berner and my family, who always have my back and give me moral support throughout my journey.

I also thank all the people who have enriched me with their specialist knowledge about medical research processes, Swiss law, pharmacovigilance, and the real-world applications of secure computation.

Contents

Chapter 1

Introduction

Secure computation techniques such as Fully Homomorphic Encryption (FHE), secure Multi-Party Computation (MPC), and Zero-Knowledge Proofs (ZKPs) allow different parties to perform computations without requiring them to reveal their inputs to each other. While these techniques have long been explored in cryptography, the last decade has seen the performance of many advance significantly, to a point where they can be used to practically solve problem instances of sizes useful to real-world applications. However, even with this dramatic performance improvement, applications of these techniques have remained relatively rare. MPC has seen a few high-profile applications [18, 17], FHE has been deployed in Microsoft's Edge browser [105], and ZKPs have been used heavily in cryptocurrencies and smart contracts [41]. However, this has so far not been accompanied by broader adoption of these technologies. Performance limitations and the complexity of deploying these techniques certainly play a role in limiting their growth. But these alone cannot account for the lack of secure-computation-based solutions even in settings where performance is sufficient and high-quality commercial solutions are available. Instead, there appears to be a significant gap between the properties and protections offered by cryptographic secure computation techniques and the challenges and concerns of those working with sensitive data in real-world settings.

In this project, we want to explore this gap in health and medical data sharing. Research on patient data is key for understanding whether and how treatments work. While clinical trials offer important insights when bringing treatments to market, biases in the composition of clinical trials can lead to issues affecting under-represented groups being overlooked [83]. In general, in today's medical research, rare diseases and side effects, off-label use of drugs, and the needs of underrepresented groups often appear only as outliers or fall through the cracks altogether. Even for severe cases, e.g. adverse reactions, it is difficult to find possible causes for the problems at hand since there is a lack of comparative material. This can be relieved by retrospective observational studies, which allow to find causation and relationships in illnesses and medical treatments in a larger scope, and thus are fundamental for the evolution of medicine. Enabling these kinds of studies is in the interest of all, but requires collecting, processing and sharing sensitive medical data, which requires solving various technical, regulatory and organizational challenges.

The majority of patient data in public health care institutions is collected for the purpose of care. The lack of a patient-owned Electronic Health Record (EHR) further leads to patient's medical history being scattered among all health care institutions they have visited. Sharing this data is a delicate matter and must be carefully considered, since patient privacy must not be violated. However, if everyone is too reluctant to share, data siloization occurs, preventing beneficial uses of the data. Due to the wide distribution of the data and incompatible or non-existent standards, even health care professionals are prevented from having a complete overview of their patient's medical history. These issues may not significantly hinder the everyday business of health care professionals (who often share data directly in mutual consultations). But this scattering and siloization runs counter to the requirements of research, where Findability, Accessibility, Interoperability and Reuse of data (FAIR Guiding Principles [123], see also §A.1.1), are considered crucial. The Swiss government criticized the fragmented landscape without interoperability, which also threatens patient safety, and the ill-incentives posed by the health insurance reimbursements as well as the siloization of data in its 2007 eHealth strategy [49]. In a 2018 update (eHealth strategy 2.0 [107]) the Swiss government acknowledged that the technological progress in the health sector lags behind other areas and that increased digital competence is needed. Digital connection and information exchange between health care institutions as well as the reuse of collected data should be fostered and supported.

A possible solution to promote data sharing is Secure Computation [69], allowing parties that do not trust each other with their data to jointly perform computations, without ever sharing their data. Today's secure computation technologies permit large-scale, highly accurate research with multiple collaborating institutions, requiring minimal trust and letting all input remain private. Using inputs from different organizations, rather than extrapolating from a local data silo, can help eliminate certain types of bias early on. This allows research to follow a rapid, iterative, hypothesis-driven approach. The unprecedented amount of data generated today enables testing of hypotheses that previously simply lacked material, such as the aforementioned rare diseases and side effects, off-label use of drugs, and the needs of underrepresented groups. However, unlocking these possibilities requires an advanced technical infrastructure, including complicated-to-deploy cryptography, and skilled professionals to operate it. Furthermore, the mental models used in creating privacy-preserving computation applications do not always match the real world when it comes to threat models as well as the legal and social hurdles that need to be overcome in order to implement such a solution. For example, such aspects are typically not considered in the synthetic data sets used to evaluate secure computation approaches [36].

To understand both the challenges and the future potential of data collection, sharing and processing in the domain of health care and medical research, we explore the current status of data handling in the Swiss health care system and investigate to what extent secure computation approaches in research and industry meet the needs of the medical community. We identify the gaps between the current capabilities offered by secure computation on one hand and the requirements of medical applications on the other hand. Finally, we provide suggestions on how to bridge these gaps.

Chapter 2

Background

2.1 Multi-Party Computation

MPC protocols allow a group of data-owners, who do not wish to disclose their data, to jointly perform a computation, where the output depends on all their private inputs [44]. In the 1980s, two-party protocols were established which allowed to solve questions such as the famous Millionaires Problem [125] (determining which of two parties has the larger input without revealing anything else about the input). Following this, multi-party protocols and protocols for more complex threat models emerged. In more recent times, computational efficiency and real-live applications moved into the center of attention. We provide a high-level overview over modern MPC approaches before considering real-world uses of MPC.

MPC protocols can be mostly divided into garbled-circuit–based and secret-sharing–based protocols. Garbled circuit protocols [51] are mainly used in a two-party setting, allowing two mistrusting parties to compute an output without the need for an independent third party. They build upon oblivious transfer and the function used in the computation is described by a (generally public) Boolean circuit [44]. In protocols for more than two parties, secret-sharing–based protocols [101] are more common. Here, data is split up into pieces, and only more than a threshold amount of pieces allow extracting the original knowledge, while having less than that amount of pieces discloses no information about the data that is missing.

In the last decade, the main focus in MPC has been to put the existing protocols to work in order to test and improve their efficiency and maturity. The two most famous projects have been the Danish sugar beet auction [18] and a large statistical study on government data in Estonia, linking tax and student data [17].

In the 2000's, a long-standing monopolist in sugar sales in Denmark had to close one of it's factories, which suddenly created the need for a nation wide auction for sugar beets. They decided to design a double auction, where the final price is computed from secret bids. Because conflicting interests made it unclear who should be the auctioneer, an MPC solution was created to relieve responsibility and prevent abuse. They used a secret-sharing protocol and assumed honest-but-curious servers in their setup. With this solution, the auction has been conducted successfully for multiple years [18].

In 2016, an MPC platform called Sharemind [15] was used to test the hypothesis of Estonian universities that working students would leave their studies early. For this, they needed to access multiple large government databases that were not linked together. Especially the tax data was protected by strong privacy laws, making such studies difficult. The team managed to successfully conduct the study with Sharemind in a privacy-preserving manner. They validated their results with a "classical" study using aggregated data in accordance with local privacy laws, and finally concluded that working while studying did not lead to early dropout [17].

2.2 Zero-Knowledge Proofs

ZKPs are often viewed as a special case of MPC with 2 parties (a prover and a verifier). However, rather than collaborating to compute a function, the prover can demonstrate knowledge of a secret without revealing it. The verifier can be sure the prover has told the truth, even when an independent observer might not be able to verify the prover's claim. We can divide ZKPs in two categories: interactive and non-interactive. In an interactive proof, the verifier interacts with the prover and thus has to vouch for the integrity of the prover towards observing parties. In a non-interactive proof, the verifier does not need to "test" the prover's knowledge in an interaction, but instead can just observe whether the prover's statement is valid [70]. ZKPs have been used heavily in cryptocurrencies and smart contracts [41], but are also used in more general-purpose secure computation protocols, e.g., when parties cannot be trusted to provide well-formed inputs.

2.3 Homomorphic Encryption

Homomorphic Encryption is the umbrella term for several encryption schemes that allow operations to be performed on a ciphertext, where the decrypted output of these operations is the same as if they were performed on the plaintext. [8] *Partially* homomorphic encryption schemes allow only one type of operation on the ciphertext, usually addition *or* multiplication. Well-known examples include the RSA [94] and ElGamal [40] cryptosystems that both allow (modular) multiplications on the ciphertext. *Fully* Homomorphic Encryption (FHE) schemes, on the other hand, allow arbitrary computations on the ciphertext. For decades, it was unclear whether FHE was even possible. In 2009, Craig Gentry first showed that FHE was feasible and follow-up work quickly resulted in first implementations. However, only in recent years has FHE become truly practical. The most recent big step for FHE was in 2017, when Cheon, Kim, Kim and Song introduced an FHE scheme that allowed approximate rather than exact values [25]. The possibility to round and approximate values made it signficantly easier to use FHE as a tool for encrypted machine learning.

2.4 Differential Privacy

Differential privacy allows us to gather information about a group without learning anything about an individual. The impact on an individual should be the same whether or not they were part of the group that was studied. This means the output of such a study is independent of the presence of individuals, thus almost equally likely to occur in any constellation of the study group. We denote "almost" with a parameter ϵ, where a smaller ϵ implies better privacy for the individual, but also a less accurate output. Note that differential privacy is not an algorithm, but a definition. There exist multiple algorithms that can compute different tasks while guaranteeing ϵ-differential privacy. Differential privacy mitigates one of the most prevalent problems of anonymized data sets, the linkage attack [38].

2.5 K-Anonymity

K-anonymity is a property of a data set which is modified such that the information of an individual entry is indistinguishable from at least $k - 1$ other individual entries. The concept was first mentioned in 1998 by Sweeney and Samariti who also suggested two methods to accomplish k-anonymity in a database: suppression and generalization [96]. Suppression means certain values of attributes, or the whole attribute column itself, are replaced by "empty" values, e.g. an asterisk. E.g., an attribute like "name", which is frequently not needed for the computations done on a data set with personal data can simply be omitted. Generalization means that certain values of attributes are replaced by a more general value, e.g., a range, or the whole attribute column is replaced by a broader category and thus all its values are also changed to a more general answer. For example, a database containing an attribute "address", where the full address of the people in the database is visible could have this column replaced by (parts of) a zip-code or even a country code, making the entry more general.

Chapter 3

Data & Digitalization in Health Care

In this chapter, we investigate how data is collected, processed and shared in the Swiss health care system. Only by properly understanding this ecosystem and its characteristics will we later be able to propose sustainable solutions.

We consider the current state of digitalization, whether there are any hindrances, what causes them and how they influence data handling. We will also explore how culture and demographics affect these processes. We observe ongoing national efforts in digitalization and complement it with lessons learned from other countries.

3.1 Data Collection

In the following, we observe the location, format and conditions under which medical data is collected. There are many different types of health data collecting instances in Switzerland and internationally. We differentiate between data collected explicitly for research and data which originates from care purposes as the respective institutions have different priorities while collecting and show varying characteristics. We review the current status and explore possibilities for change where it is necessary.

3.1.1 Collection for Explicit Research Purposes

Data collection specifically for research purposes usually happens in the context of clinical trials or through supervising bodies like SwissMedic [117] who watch for critical incidents (e.g., adverse drug reactions) that would require further investigation. We found that data collected for research purposes is usually digital and well structured, as this has a naturally high priority for the organizations collecting it.

Supervising bodies usually do not have to actively find the affected patients since there is an obligation to report [116] unexpected severe incidents in connection with medical treatment. Because doctors are obliged to also include relevant parts of a patient's medical history, the incidents can then be investigated by national research centers. In these reports, a patient's identity is anonymized, and thus no consent from the patient to disclose the information is needed [9].

Finding data for a clinical trial, on the other hand, is time-consuming, expensive and slow. These projects need to be approved by an ethics committee and suitable people for the trial have to be identified without compromising data protection laws [72]. In many countries, there exist registries where people can volunteer to be contacted for trials [12], but this is still relatively rare in Switzerland. There are also various organizations trying to build such databases all over the world to facilitate research on new treatments. However, there are often inherent biases in clinical trial groups, as clinical trials have to adhere to strict regulations regarding patient safety and companies do not want people to drop out of the expensive process. As a result (pregnant) women, the elderly and demographic minorities are often excluded, leading to issues specific to them being overlooked [83].

The expenditures of clinical trials lead to the question of reusability of the data, since it might be more cost-efficient to work with existing data. According to the law, health-related data taken for research purposes may be disclosed for research purposes, which would encourage reuse. [112] However, the rules of data proportionality and data minimization also apply. This means that data cannot be accumulated arbitrarily simply because it might be useful at some point in the future. Therefore, the collected data is likely of highly specific nature, such that it is rather improbable that it would be useful for another type of study.

3.1.2 Data collected for Care and Treatment

The vast majority of medical data arises from the health care environment. Even though it comes in varying quality and levels of detail, this type of data is interesting because of its broad scope. Data collected in a care environment is comparatively unbiased as it includes everyone visiting a doctor, making it ideal for research. For example, it could be used for large-scale retrospective studies and as input for machine learning, which would allow one to, e.g., discover drug interactions, assess treatment feasibility over extended periods of time or for demographic groups not considered in the clinical trial, or detect early indications for certain diseases. However, there can be other forms of bias, such as due to doctors' and hospitals' social and demographic catchment areas. While health insurance is mandatory in Switzerland, there is an extensive range of offers and some treatments are only covered by expensive optional insurances. This means that some providers might be mainly visited by wealthy patients, while others (e.g., focusing on general or emergency care) see a wider variety of patients.

Most medical care is provided by general practitioners, making them an essential potential data source. However, in Switzerland there are no national standards prescribing how to collect patient data. As a result, some practices are still fully paper-based while others use a variety of different electronic systems. Some take part in the FIRE project, which aims to build a nation-wide database of family medicine, with data from routine check-ups as well as diagnoses, classified using the International Classification of Primary Care (ICPC) standard. However, electronic or not, patient records commonly consist primarily of free-form text, making them ill-suited for most automated analysis. Hospitals' patient data tends to be better documented and structured. However, the files still suffer from weaknesses of the individual documentation tools and frequently do not follow the FAIR principles [123].

The federated landscape of Switzerland makes it hard to define consistent codes and standards. Even if everyone could agree on the same codes, these would need to exist in at least four languages. If the codes are poorly translated (e.g., initial versions of Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), see §A.1.2), information is lost and they will likely be abandoned. Since the law lacks an enforcement of a standard, the main working incentive for adapting a uniform standard are the health insurances. They can decide in what form they accept reimbursement requests [81], which is the reason that hospitals mainly use the International Classification of Diseases (ICD)-10, Swiss Operation Classification (CHOP), or SNOMED CT code systems (see §A.1.2). While this means health data from insurances is very well structured, it might still be biased due to conflicting interests and incentives. Health insurances in Switzerland employ a fixed-price system for every type of diagnosis and intervention, which does not take into consideration the individual circumstances, including the length of the hospital stay. Therefore, there is an incentive to identify as many diagnoses as possible, especially in patients requiring more time and attention, resulting in a distortion of data sets.

3.2 Data Sharing & Processing

In this section, we explore existing data sharing and processing frameworks. We will be investigating how access to collected medical data is controlled in practice before assessing the legal restrictions on this sensitive kind of data. In the context of research, the most important properties of data are its shareability and interoperability. Therefore, we pay special attention to whether or not existing sharing processes are suitable for medical research. Shared data can also suffer from significant information loss through aggregation techniques used to adhere to data protection laws. These factors need to be considered carefully, because they might render the data useless for projects that require a high level of detail.

3.2.1 Accessing Collected Medical Data

How and under what circumstances data collections can be accessed differs from institution to institution. We found that trust among different parties in the medical domain is very high in Switzerland, and exchanging medical data between institutions for care purposes can be surprisingly easy. Generally, neither the research community nor the collecting institutions seem to consider malicious actors. However, most processes do require the researchers to authenticate themselves, and some even require having direct contact in the data-sharing institution, who might carry out the analysis themselves rather than hand out the underlying data.

Most governments maintain incident databases for analytical purposes and surveillance of patient safety [75]. For example, SwissMedic has a National Pharmacovigilance Centre and they periodically crawl their database for quantitative signals that indicate that a drug needs a closer inspection. While this data does not seem to be accessible to independent researchers, we found that they are often monitoring the Direct Health Care Professional Communications (DHPCs) where SwissMedic communicates their findings and might conduct independent studies on the drugs in question.

Some Swiss hospitals also allow researchers to use their data in the form they have the patient's consent to do so [11]. In the private sector, some health insurances make anonymized data commercially available to researchers. We found that data analysis is usually done in-house for data protection reasons, with no data leaving the insurance [118]. Several pharmaceutical companies, like Novartis and Bayer, also share some of their data with researchers, e.g., via a collective portal, where a review panel evaluates research proposals and grants access to patient-level data [28]. This data tends to be well organized and documented since clinical trials by pharmaceutical companies are highly regulated [3]. Non-profit organizations in the medical field often actively support research as well, e.g., the Gates Foundation with its Gates Open Research project [47] provides funding to research resulting in public data sets. If the data is too sensitive to be published, researchers can define access constraints such that the foundation can handle access requests. Other organizations also offer data upon request or combined with industry collaborations, e.g., Médecins Sans Frontières [78] and the Institut Pasteur [64].

More convenient data sources for researchers are being built by the Swiss Personalized Health Network (SPHN) [113] that is funded by the Swiss government and tasked with increasing health data availability for research. Other governments, for example the U.S., already have organizations that allow statistical queries on some of their databases [22]. The Centers for Disease Control and Prevention (CDC) has an interactive database system that allows querying e.g. the number of cancer patients according to the year they were diagnosed, the local registry they were reported to, the body part their cancer was found on and the age range they were in when they got the diagnosis. If the database finds less than 16 entries for a query, it does not display the results. The CDC further has a research data center that gives access to restricted-use data to authorized persons. The World Health Organization (WHO) gathers data from all over the world and makes it available for research in different ways [24]. In aggregated, anonymized form, the data is available for national research centers. For specialized queries, it also appears possible to informally receive more detailed data, if one has contacts to the WHO and the necessary references.

The lack of standardization makes most of the data gathered by general practitioners unavailable for research, even though it makes up a huge part of medical data in

Switzerland. However, currently, more than 700 doctors in Switzerland already take part in the FIRE project [26, 31], making some care data available for research. To use FIRE project data for research, it is anonymized by only transmitting birth year, gender and the patient's case number as potentially identifying attributes. The Swiss data protection agency has approved this procedure. This means the data can be used for research without explicit patient consent. However, the data base can only be evaluated by the scientific staff of the institute for family medicine at the University of Zurich. Which projects are conducted is decided by a committee of university and family medicine representatives. [45]

Finally, we observe that health care institutions tend not to actively share data amongst each other for care and treatment purposes. Instead, health care professionals can exchange (anonymized) thoughts in specialized networks [119, 56] and urgent information requirements are resolved via calls or emails between different institutions. While many are using the secure services from Health Info Net AG (HIN) [53] for data exchange, it still happens that paper-based records are faxed or scanned and emailed without encryption. For example, in the early stages of the Covid-19 pandemic, positive test results had to be entered in an electronic form, which was then *printed and faxed* to the Federal Office for Public Health (FOPH) [20]. This culture of low-threshold communication is non-standardized and seems to be based on high mutual trust, but appears to work well in practice. However, in some cases the ad-hoc and opt-in nature of these channels might lead to important information being lost.

3.2.2 Legal & Regulatory Considerations

To fully understand the constraints under which this data can be shared and processed, we need to take a closer look at the legal frameworks in effect. In Switzerland, cantonal ethics committees [84] are generally responsible for approving research projects involving human subjects, such as medical research, as well as for approving exceptions permitted by law. [112] The committees have to ensure that laws are obeyed and guarantee study participant's safety, dignity and well-being. This also holds for retrospective studies and thus for research on data collections [68]. Here, we highlight the impact of regulations like the Federal Act on Data Protection (FADP) and Human Research Act (HRA) on medical research by investigating the prerequisites for data processing under Swiss law.

Consent & Alternatives If data will be used for research purposes at some point, consent has to be obtained from the affected person "in the case of processing of sensitive personal data or personality profiles" [110]. The HRA specifies that research on humans is only permitted if the participant has been sufficiently informed and consented to only a specific procedure, which is known as *informed consent*. A patient can be asked for general consent, making their data available for current and future research projects. In certain cases, the use of the data is considered particularly dangerous, and general consent is not sufficient. This applies, for example, to the further use of biological material or genetic data.

On the other hand, the law also provides ways for research to occur without explicit consent. If (non-genetic) personal data is used in de-identified form, the data subject must be informed of this, but consent is not required. If the data are anonymized, their use for research is even permitted without any preconditions. The “research privileges” in the HRA come into play when data is processed in research for non-personal purposes, anonymized during processing as soon as the purpose permits, and the persons concerned cannot be identified in the publication of the results. Then the need for consent does not apply, even if the processing of the data was previously explicitly rejected by the person. [112] The Message on the FADP also states that a commission of experts appointed by the Federal Council may authorize researchers to view medical records without patient consent if it is impossible or disproportionately difficult to obtain this consent from the patient in question, the study cannot be conducted on anonymized data and if the research interests outweigh the secrecy interests of an individual. This is meant to balance the public interest in the advances of medicine and data protection rights. [111]

Identifiability of Records Many legal and regulatory frameworks provide exceptions for “anonymized” data. However, medical records usually contain data that makes a person easily identifiable. To remove identifiability (known as de-identification), we can either *anonymize* data, which irreversibly removes the link between a person and their data, or *pseudonymize* it, where the connection between a person and their data is stored somewhere, rendering the action reversible. [77]

De-identifying data is a challenging task. Especially in data files containing free text, it is extremely hard to automatically remove all identifiers from the free text, while doing so manually will not be feasible for many projects. Even beyond the challenges of dealing with free-form text, there seems to be no uniform approach to the de-identification of medical records. Especially for small institutions without a legal department, the uncertainty resulting from the imprecise formulation of the data protection laws might discourage sharing efforts. But even large institutions dislike anonymization because it removes valuable information, sometimes to the point of rendering data useless altogether. Instead, larger organizations prefer to use pseudonymization, which leads to siloization as the linking information cannot be shared.

The processing of anonymized and pseudonymized data (for those that cannot re-identify the subject) is not covered by the FADP because the data is no longer considered personal data. While convenient, anonymization is sometimes impossible. For example, the (message to the) HRA states that in research projects related to severe diseases, the data collected may not be anonymized so that researchers can contact and inform the patient concerned if necessary unless the patient has expressly waived information. Also, in long-term studies, the relevant data cannot be anonymized in most cases since the link between the person and the data must be restored on an ongoing basis. [112]

3.3 National and International Efforts in Digitalization

In this section, we give an overview of the digitalization efforts in the Swiss health sector and compare it to international counterparts. We will first consider Switzerland's characteristics and current direction before we describe the situation in other countries, especially in the U.S. and Estonia. The latter is especially interesting for us because of its highly advanced digitalization that runs through all parts of its governmental bodies.

3.3.1 Digitalization of the Swiss Health Sector

Federalism has always defined the Swiss health system. Some institutes, such as private hospitals, operate under federal law, while other universities and public hospitals operate under cantonal law. Further, we can see that FADP as well as General Data Protection Regulation (GDPR) offer little guidance for security and medical professionals alike, who have to choose their privacy-preservation measures carefully. This fosters an aversion towards technological advances since they first would need to be assessed under these obligations. When it comes to standardization, an essential pillar of digitalization, we need to consider Swiss multilinguality, which makes standardization an unusually arduous task on a technical level. When considering the status quo, an obviously missing step towards digitalization and especially automation is the ability to handle data at scale. Many operations on medical files are conducted manually, which heavily limits the possibilities to scale these processes.

The 2018 eHealth strategy indirectly admits that many of the original problems mentioned in the first strategy [49] from 2007 still persist. As a result, the SPHN was founded. SPHN is lead by the Swiss Academy of Medical Sciences (SAMS) and collaborates with the Swiss Institute of Bioinformatics (SIB), the Personalized Health and Related Technologies (PHRT) of ETH and interacts with international data-sharing initiatives "to ensure lessons are learned". It is tasked with solving the major pain points in medical digitalization, namely build marketable coordinated data infrastructure to share interoperable health-data for research in Switzerland. SPHN is trying to bridge the interoperability gap between competing standards by adding another layer of indirection and thus accepting all previously used standards in its data model with its semantic interoperability framework [19]. Today, SPHN has 24 driver and infrastructure development projects [89, 104]. One of them is DeID [27], a project that should clarify how medical documents should be properly de-identified, with the goal of an implementation that works in German, Italian and French. SPHN further invests in Natural Language Processing (NLP) projects to transform free text into known terminology and connects cohort data to a metadata catalog for findability. NLPforTC [4] wants to build an NLP-based tool that maps clinical reports to SNOMED CT codes.

Another project deals with e-consent [95], developing a "harmonized interactive electronic general consent" to make the process of (informed) consenting easier, faster and more broadly applied. We observe that the progress made thanks to SPHN is impressive when it comes to possible solutions to many issues named before. The challenge of nationwide adaption of these solutions remains.

The basis of a secure infrastructure, called BioMedIT, have been set up by SIB. It should allow authorized researchers access to confidential data with the goal of collaborative analysis. Of course, data privacy is highly valued in this setting. The three main nodes are now operational in Basel, Lausanne and Zurich. The data that is uploaded in encrypted form by health care institutions to a node is shared with the other nodes. Authorized researchers can access them through a single portal with two-factor authentication (2FA). The project that might be closest to our topic is MedCo [74]. MedCo aims to develop an open-source, privacy-preserving operational system for health care institutions that makes their data available for research in a secure, distributed way, using homomorphic encryption, secure MPC and result obfuscation. MedCo plans to offer secure, private cohort exploration. It already supports the existing i2b2 framework, which should make it easy to deploy it on top of i2b2-based infrastructure. They even promise a similar query response time for cohort exploration as the non-privacy-preserving i2b2. However, it was started as a research project and SPHN is still in the process of making MedCo production-ready.

In summary, it seems Switzerland is lagging behind its potential when it comes to digitalization in the health sector. However, we welcome the efforts as part of the new eHealth Strategy, which already show promising prospects and give reason to believe that Switzerland will be catching up as long as government support is maintained and digitalization stays a national priority [29, 79].

3.3.2 International Comparison

In the following, we will review some examples of digitalization in health care from other countries and thus allow a comparison to the situation in Switzerland. We will focus on Estonia and the U.S. because Estonia has a technically highly advanced infrastructure and the U.S. has seen a boom in tech startups over the last few years thanks to Health Insurance Portability and Accountability Act (HIPAA) [73].

Estonia recognized that establishing trust is something that needs time, and that this also holds for sustainable digital progress. They found it essential to be transparent at every step of the way, build informal types of communication to the citizens that lower the entry barrier, and let systems run internally for a while before they are opened to the public to eradicate technical teething troubles [85]. The digitalization process was promoted as an anti-corruption measure as it frequently removed "human" steps from official processes. Since 2018, whenever officials process a citizen's data, the citizen has to be informed about it and provided with a contact possibility (with the exemption of criminal prosecution) [67]. Most e-government solutions in Estonia are open source for transparency [42]. The government has also nominated digital advisors whose job it is to bring the technologies to the people, e.g. through hold digital conferences where the audience can directly ask questions about the different projects [39]. Technically, Estonia's e-government is built around their e-identification and national public key infrastructure, allowing citizens to authenticate themselves for public services. Another key ingredient is X-Road, an open-source, decentralized ecosystem for data exchange that connects all public and some private institutions, using a blockchain solution to log events in order to prevent fraud [67]. The government applications are designed for ease of use, and use decentralized systems. Thanks to their already advanced IT infrastructure, Estonia managed teleworking, telemedicine and teleschooling during the Covid-19 pandemic much better than other countries [35]. It is worth noting that Estonia works mainly with local companies since at the time, Estonia did not have the financial means to pay big, established companies. As a result, the private sector and the government have worked very closely and there is a significant amount of rotation of employees between the government and the private sector.

While there may be quite a number of cultural and demographic differences between Estonia and Switzerland, we can still learn a lot from their approach. The success of establishing digitalization as a part of a countries DNA, making highly complex technologies in short accessible to a broad audience, introducing strong legal protections, and adoption by 99 percent of the population did not come overnight. It took uncompromising transparency, the will to collaborate cross-sectional, and broad inclusion of citizens, academics, government institutions and private companies, with the goal to build trust at the center. An example that could maybe be reproduced in Switzerland is how the e-health system in Estonia allowed the country to establish several registers for rare diseases, a field where research data is usually scarce [91]. Legally, this would already be possible in Switzerland, since the creation of such registers is not in itself research. Thus the HRA does not apply, and no consent is required at the point of creation, as the Swiss federal ethics committee has clarified [46]. Not only in Estonia, but also in the U.S., technologies like differential privacy or MPC are slowly making their way into government processes. Since 2020, the U.S. Census Bureau has been using differential privacy to enumerate people and households of U.S. citizens, after it has previously used the less secure imputation that likely allowed linkage attacks [34]. In 2017, a bill was introduced to U.S. Congress called ”Student Right to Know Before You Go Act”, which should make results available about student's graduation rates, debt levels, salary and other units of different universities. The goal is to properly inform students how to spend their college fund best. Since these measures can only be calculated with sensitive data from different federal offices like the Internal Revenue Service and the Department of Education, they suggested using MPC to combine the data. While the bill is still being considered in Congress [127, 124], it shows that legislators are aware enough of secure computation to identify suitable deployment opportunities. Clearly, there is some room for improvement in Switzerland in comparison to other countries. However, the long-established political and economic stability in Switzerland does not afford it the same "fresh start" scenario that enabled Estonia's drive to digitization. Rather than looking for a huge leap of progress in a short time, we need to consider the more profound lessons of trust and transparency and how they can apply to Switzerland.

Chapter 4

Secure Computation in Practice

In this chapter, we explore current (commercially available) secure computation approaches and how they are or could be applied to the needs of the medical community. In addition to secure computation platform providers, we also consider companies offering security compliance and policy tools, as such solutions are currently more widely adopted.Finally, we will assess the current approaches to secure computation in research and how academic research is being brought to market. Instead of focusing on the technical implementation of different solutions, we focus on how well they address the requirements of real-world medical applications.

4.1 Commercial Secure Computation Solutions

We explore the growing number of commercially available secure computation solutions, analyzing how suitable they are for data analytics on health care data. Our survey is not exhaustive, but covers examples of technologies and companies we found most ready to ship and most frequently mentioned in the industry. We focus primarily on MPC, where commercial offerings seem most mature, but also briefly address FHE and other secure computation technologies. Several companies offer ready-made platforms for secure MPC, the most well-known of which is Sharemind by Cybernetica [102, 15]. These platforms for privacy-preserving computations are meant to facilitate the process of setting up the software part of such an infrastructure. Next to Sharemind, we will consider the smaller, younger company Inpher as a representative for the emerging startups, and compare their offers.

Cybernetica's Sharemind MPC framework promises computation on encrypted data with comparably low performance overhead, which can be hosted in any data center or cloud as long as at least three servers can be used as nodes. They further enable external privacy controls that manage what computations and outputs are allowed, such that no information is leaked. Private values are stored in a secret-shared way, such that no single server can learn them [98]. There are a number of papers based on Sharemind, e.g., linking (simulated) health records from different health centers together if they belong to the same person to avoid duplication in a privacy-preserving manner. With Rmind, Cybernetica also designed a cryptographically secure statistical analysis tool based on the R language [16] to make statisticians feel more at home when working with MPC. With the help of Rmind and Sharemind they conducted a well-known real-world MPC study (”Students and Taxes” [17]) demonstrating that their offerings are ready for real-world use, linking and analyzing ten million tax records and 500’000 education records without loss of accuracy or privacy. The results were validated with a classic statistical analysis, which had to use anonymization methods to comply with privacy regulations, leading to a significant sample loss of 10-30 percent, while MPC suffered none. The bias introduced by k-anonymity aggregation further lead to 4-13 percent difference between the anonymized and the MPC-based precise results.

Inpher [62], a Swiss-American Startup, also develops a secure MPC solution called the XOR Secret Computing Engine. Inpher claims to provide high accuracy and precision and compliance with cross-border data transfer standards [1]. In 2020, Inpher won 2 of 3 tracks in the iDASH secure computation competition, exceeding prior state-of-the-art performance results [60, 63, 61]. With Manticore [21], they also propose an MPC framework specifically for differentially private federated learning. Manticore is designed for real number and Boolean arithmetic as well as garbled circuit operations, which include real-valued polynomials, division, exponential, logarithm, linear combination, and (oblivious) comparison and requires a trusted dealer. Several applications are possible with this range of operations, and the designing team focused on logistic regression, PCA and oblivious sorting. However, as far as we know, there are as of yet no real-world deployments of either Inpher’s XOR or Manticore technologies.

We see that these companies already have medical use cases on their radar and show highly promising results. The benefits of such offers are that data analysts can implement queries without having access to the data and only the issuer of a query can read the output, offering high security and confidentiality. Furthermore, they promote new forms of collaboration and previously unknown scopes of data, which allow to investigate hypotheses that were impossible to test before The main drawbacks of these companies are that often a trusted third party is required, as well as a technically highly advanced infrastructure and skilled professionals to handle the complex setup. Even given a performant infrastructure, most queries will take a significant amount of time, but recent improvements promise to bring down the query time in the future significantly. Further, each question has to be implemented and solved individually and one might need to organize a party supervising information leakage to ensure privacy, creating a vast setup overhead. This is a skillset that, e.g., researchers in Swiss hospitals will not have at hand in order to conduct cross-institutional studies. Therefore, these technologies will have to become significantly more accessible before they can be deployed at scale.

4.2 Compliance Tools

Currently, most security and privacy efforts in industry concentrate on encrypting data in transit and at rest, and use non-cryptographic techniques like access control to control processing. As a result, compliance, policies and out-of-the-box security solutions are mainly built to keep the data where it belongs, rather than to share it securely, furthering siloization [86]. A variety of companies realizing and selling a more compliance-centric vision of privacy-preserving technology exists, focusing on high-quality access control, privacy policies and heuristic measures such as de-identification.

Privitar [87], for example, is a security company offering many services for de-identification and linkage of data automatically and at scale, access control, encryption, implementation of privacy policies, cloud security, watermarks and monitoring. The techniques they use for de-identification include pseudonymization, generalization, perturbation, noise addition and data minimization. A so-called Protected Data Domain maintains the structure of the original data set, keeping format, storage location and meta-data, which eliminates the need for adapting existing applications that work with this data [88]. One of their use cases is health data protection and analysis. They have partnered with the National Health Service (NHS) where they have mitigated the issues of federation by linking the data from the data silos together in a (presumably) secure and private way with the help of homomorphic encryption. [57] Since the NHS did not have the technical infrastructure to handle such a large amount of data, they started to use the Amazon Web Services (AWS) cloud services. [120]

Of course, companies like Microsoft and Amazon also offer their own compliance tools. AWS provides an extensive list of services for Identity and Access Management (IAM), threat detection, infrastructure protection, data protection, incident response and compliance [5]. Creating custom solutions by navigating this abundance of possibilities tends to be cumbersome, making it error-prone. As a result, a significant amount of data breaches are enabled by misconfigured access control systems. Meanwhile, the Microsoft Security Compliance Toolkit [76] allows one to edit and analyze security configuration baselines for Windows clients and servers. The tool shows security administrators where and how these baselines should be applied. However, these cover only the bare minimum of security, and do not consider the customizations done by individual companies. There, the security administrator has to decide themself where to add more rules. Tools like this are becoming easier to deploy, but health care institutions have to make sure that all their assets are truly covered by them and also adhere to the local law. While helpful, such tools can also lead to a dangerous sense of perceived security, leading to complacency rather than compliance.

4.3 Real-World Example: Contact-Tracing Apps

While the progression from academic research to real-world ready solutions has traditionally been slow in this domain the Covid-19 pandemic has seen the rapid deployment of privacy-preserving contact-tracing apps in many countries. Interestingly, some of these use complex privacy-preserving systems based on academic ideas, while others rely on more traditional and privacy-invasive approaches. In this section, we investigate the technology behind these two fundamentally different types of contact tracing and set them in context of their public perception and adoption. This gives us information about possible pitfalls that privacy-preserving systems need to manage if they are widely deployed.

The Decentralized Privacy-Preserving Proximity Tracing (DP-3T) [37] protocol has been the base for several contact tracing apps in Europe, namely Austria, Belgium, Croatia, Germany, Italy, Ireland, the UK, the Netherlands and Switzerland [10]. The technology uses Bluetooth Low Energy to track contacts in the background without draining much battery, an approach that had to be supported by the phone's operating system (OS) providers Apple and Google in order to work [50]. The vital core of the protocol is the so-called Ephemeral IDs used as an identifier. The strings are 16 bytes long, semi-random, rotate in specific intervals, and are used for logging the contacts of the person using the app. The solution offers strong privacy guarantees and is a rare example of an academic design being brought to market in record time. In July 2020, about 14 percent of the Swiss and German population had downloaded the app, while in Italy, it was only 7 percent [114, 82, 32].

Meanwhile, in South Korea, researchers developed a Privacy-Oriented Technique for COVID-19 Contact Tracing (PROTECT) that uses homomorphic encryption to share the location of patients and a secure proximity computation and allows a central authority to notify people if they are within 100 meters of a patient [6]. However, instead of adopting this protocol, South Korea chose to release the (somewhat de-identified) travel histories of confirmed patients and displaying coronavirus-hit areas. Gender and age range of patients seems to be disclosed in addition to the travel histories, exposing patients to a high risk of re-identification. Despite these possibly massive privacy violations, the acceptance of these measures appears high [126]. While South Korea has made its experiences with the Middle East respiratory syndrome (MERS) in 2015 [71] and thus knows of the necessity of countermeasures in a pandemic, there are also cultural factors that influence the acceptance of such actions.

It appears that the trust in authorities, the health care system and technology is frequently more important than the actual quality of privacy-preserving designs. In Europe, concerns like fear of government surveillance and fear of being stigmatized have been a significant barrier to adoption. Additional influencing factors are the individual's perception of social responsibility and technology, as well as the perceived threat to their own health. The individual's understanding of the functionality and preventative nature of the app, as well as what performance and benefits can be expected from it play a crucial role in broad adoption, as the spreading of misconceptions can significantly hinder the process. Last but not least, we need to consider an intention-action gap, where people have a positive perception of the application but still remain passive [121]. We observe that this displays again the crucial role of transparency and low-threshold educational work in technological competence, especially in Europe.

Chapter 5

Discussion: Bridging the Gap

In this chapter, we will think of the measures and changes required to close the gaps between secure computation offers and real-world health care and medical research requirements. We divide this into matters concerning ease of use, the public perception and their security concerns and the underlying protocols and threat models of secure computation technologies. We will discuss possible solutions and how a pragmatic attitude in research and client-centered product development could go a long way.

5.1 Ease of Use

In the following, we discuss the usability shortcomings of secure computation offers using the example of MPC. We suggest ways how ease of use could be improved generally, but specifically in the context of medical research.

As of today, MPC applications are usually custom-tailored for specific applications. This effort is hardly scalable since applications require an elaborate set up and a lot of preparation work for query and data set preparation. Not all projects have so much time to spare or the courage to tackle such a technologically demanding task and thus might rather keep working on smaller, centralized data sets. It should be a goal to get away from highly customized systems to more standardized and maybe even certified approaches that offer certain legal guarantees and robustness to customers. Ideally, off-the-shelf systems will reduce the preparation time by orders of magnitude, such that willingness to experiment with this technology increases.

Not every project is suited for MPC, as some operations will remain difficult to perform on encrypted data (e.g., NLP) or will lose information in the process of preparation and encryption that would have been necessary to answer the research question. Researchers considering MPC should first explore whether or not their techniques translate to MPC, as well as whether they can actually benefit from an increased sensitivity and scope of the data because of MPC. Lastly, they need to consider whether the question they want to answer can be formulated as an MPC query without leaking information unintentionally. Companies offering MPC platforms should provide support for researchers to clarify these questions to avoid frustration with their product or accidental information leakage. We identify here a potential business field where companies could sell solutions for specific problems, thus providing ready-made, off-the-shelf setups. For example, an offer for private set intersection on patient data containing a particular list of value types and identifiers, linking a fixed amount of parties and requiring at least a specific size of the input data sets. These offers could reduce the know-how necessary from the client and thus increase their catchment area. Modular designs with frequent components such as set intersection or duplicate removal also offer good reusability for the service provider.

Further, ease of use can be massively increased by considering the clients who will use it later. For example, Sharemind developed the Rmind tool that is modeled after the programming language R frequently used by statisticians [16]. Providers of cryptographic solutions must first understand the tools people already use and integrate their product into this environment. It is also more constructive to work with the habits of people and not against them, which means e.g. parsing free text with NLP instead of requiring people to enter data in a highly complex fashion.

We assume that work on unencrypted data will also remain necessary since we saw that not every research question can be put into an MPC query. This type of research scales poorly today because we found that time-consuming tasks (e.g. assessing and performing de-identification) are frequently done manually. For this reason, we consider it crucial that repetitive, question-independent tasks like anonymization become automated. Some companies like Privitar already offer such services, but as the medical world is little digitalized and standardized, it remains a tedious task which we imagine needs the support of NLP to be truly performant. This does not necessarily solve problems like linkage attacks, but applying differential privacy could mitigate this while retaining high accuracy, and various tools and libraries for differentially private machine learning have become available over the last few years [52].

In medical research, linking data sets is sometimes actually a necessary task because of the scattered medical histories of patients. Linking of course needs to be considered before the data is fully anonymized, and also requires careful consideration of identification possibilities [43]. One possible approach could be to hash or otherwise encrypt specific identifiers before removing them, such that matching identifiers can be combined cross-institutionally. This of course requires that the combined data still does not leak information about the patient's identities, which is inherently difficult to know before the data is combined. This consideration makes it more attractive to use something like MPC, where the encrypted inputs do not necessarily require anonymization, and only the output of an analysis is potentially public. There is still a requirement for a carefully crafted query, but it could mitigate some of the issue.

We can summarize the suggestions above as a call to client-centered product development and an invitation to orient oneself on already commercially available security products and the way they are presented and sold.

5.2 Public Perception and Security Concerns

Secure computation is not a widely known solution for data-sharing in Switzerland and the legal and societal frameworks have not yet been adapted to it, leaving hurdles regarding consent, ethics and the law. These obstacles make even the existing offers unattractive as many companies will not see an advantage over the restrictions of classical data-sharing options. As we have seen in the example of Estonia in the previous chapter, public perception can have a great influence in overcoming such obstacles. Public perception plays an even more significant role in the adoption rate of new technologies, as we have seen in the example of contact-tracing apps. The government support for the Swiss Personalized Health Network (SPHN) demonstrates that the shortcomings are known, but we believe it is crucial to increase the visibility of their work and explain their necessity and prospects to the public to push a broad adoption when the products are ready.

While secure computation promises new ways of collaboration and research on sensitive data, it can not mitigate all human issues in security compliance. There are many reasons why security programs can fail [86], including misjudgment of which assets are at risk, or a lack of inclusion of the companies specific structure. Combined with an aversion to deal with the complexity of security, decision-makers are tempted to buy out-of-the-box solutions or outsource the responsibility, potentially leaving the most important assets insufficiently secured. The same aversion might also lead to a lack of verification of whether security mechanisms are correctly installed and maintained. Since the responsibility for security is often attributed to the IT department, decision-makers might be tempted to focus mainly on the technologies, instead of looking at the whole picture of the organization, its assets and risks. Homegrown security solutions might fall victim to the same issue as long as there is not enough investment in building up the necessary know-how and responsibility and execution are not sitting at the same table. Another reason for security to fail can be an over-focus on legal compliance. Especially in hospitals, there is a strong focus on compliance with applicable laws and regulations. In fact, some hospitals have their own well-crafted guidelines that are even stronger than what the law requires. However, the law does not mention all the numerous side-channels hackers can take to get what they want, leaving an organization vulnerable even though they implemented everything they thought they needed. There is no way around the classic security concerns like access controls, policies and compliance. Since every organization's resources are finite, it is imperative that a thorough threat assessment is conducted and security mechanisms are applied accordingly, fit to the specific organization's need with the help of professionals.

Security and its maintenance are and will always be a complex topic, which is why trained professionals are important [86]. We believe this also holds for the deployment of secure computation solutions since they often require on-site processing of sensitive data, for example to encrypt them. Even cryptographically secure tools are still used by humans and could be abused by unauthorized personnel or infected by malware. This is why secure computation companies should consider extending their offerings in the direction of traditional compliance frameworks. Such a push could offer a more gradual path to advanced security technologies rather than forcing an all-or-nothing decision. This could help relieve the initial aversion we frequently observe in general audiences because of the high complexity of cryptographic solutions.

5.3 Protocols and Threat Models

In this section, we will investigate what needs to change in secure computation research in order to be a better fit for real-world and specifically health care applications. Current data sharing efforts often require a lot of trust and paperwork, as well as a considerable number of resources in a centralized data store, like the previously mentioned research databases. MPC promises to mitigate these challenges by omitting (at least partially) the trusted party and keeping the inputs private, as well as distributing the amount of resources needed over the participants. We could avoid data silos, legal issues and country borders. This idea has paved the way for specialized companies as we see currently in very early stages, and quite advanced research areas. However, the offerings of these companies and researchers very often still rely on trusted third parties [62] or can only offer 2-party-protocols [36], which might not bring enough benefits for clients. Since clients will invest many resources in such a solution, we can assume that they will not simply want to merge their silo with another but significantly scale their coverage. The legal implications of a trusted third party are also a delicate matter, especially in medical research.

Furthermore, we witness that the threat models used by cryptographers can be an unnecessarily hard restraint when used in real-world applications. Most institutions do not have to care about computationally unbounded enemies nor do they consider scenarios where they can trust no-one except themselves. While critical infrastructure should ideally be safe enough to keep out even nation-state actors, we believe such threat models are too much of a limitation for medical research applications. Attackers with such capabilities can usually find significantly easier ways to subvert an academic research project than running infinite queries on an MPC setup. We thus believe it is okay to trade, e.g., potential leakage in the case of infinite queries for efficiency.

We found that threat actors stronger than honest-but-curious do often not match the participant's perception of reality. In most cases, all parties are trusted, and MPC is only used as a data protection measure to adhere to the law and not because the parties would not trust each other. In the field of medicine, the trust is exceptionally high, as they have to prove themselves over and over in handling sensitive data. This does not mean that we should abandon threat models in MPC, on the contrary, we should simply consider a different view. We usually assume the contributing parties to each be a single entity. This is not necessarily true. If we come back to our example of hospitals, from collecting the data to feeding it into an MPC algorithm, there might be dozens of people involved, some even accessing the same computers. We do not believe that all of these sub-entities will always act like trusted parties. A marketable threat model therefore needs to include a certain robustness when it comes to subentities of a computing party.

Based on these insights, we consider two types of threat actors that are generally not considered in secure computation but of significant importance in practice: A *byzantine-by-incompetence* actor is (at least somewhat) technically illiterate. They do not know what they are doing, turning computers and programs on and off and randomly clicking on stuff they are not supposed to. Moreover, they might enter data in the wrong order or load the wrong files. While this type of byzantine behavior is technically captured in the actively malicious threat model, they have little in common in practice. A byzantine-by-incompetence actor can be mostly be mitigated by employee training, access control, graceful shutdowns and general failure-tolerant designs. However, few of the existing secure computation systems seem to consider such users or provide the facilities to effectively deal with them. Our second actor, the honest-but-incentive-driven actor technically follows the (MPC) protocol correctly but has a different agenda than whoever set up the computation. They manipulate the input to be valid, but not necessarily correct. They might selectively decide which data to collect or exaggerate on parts of the data, resulting in biased and distorted inputs. A more direct form of this adversary would simply manipulate the inputs directly to hide, falsify or add inputs in accordance with their agenda. Such actors tend to arise when systems are used for multiple, misaligned purposes (e.g., recording medical diagnoses and managing remuneration). While they can occur in traditional systems, the reduced auditability of privacy-preserving systems makes them more likely. A honest-but-incentive-driven actor is harder to mitigate, requiring good knowledge of the systems around this actor, their motivation and roles. Ideally, misaligned incentives are detected and discussed before a system is set up. Alternatively, performing statistical analysis on different (aggregated) data sets from similar sources might also reveal certain biases, but only if it is not an industry-wide problem.

Chapter 6

Conclusion

The goal of this thesis was to explore how medical data is collected, shared and processed in care as well as research, analyze how secure computation could solve the needs that arise from these procedures and find what is required to close remaining gaps between secure computation offers and real-world health care and medical research requirements. We used the example of health care because there are many application scenarios with obvious societal benefits. It was also interesting because it has seen very little exposure to advanced cryptographic technologies in comparison to, e.g., the financial sector.

Our analysis of the current data handling in medical research has revealed several significant challenges, primarily related to (a lack of) digitalization. When exploring the fit between the medical and secure research communities, we found that several exciting prospects exist, for example the SPHN tying technology and the medical field together to develop MedCo, a secure operational system for hospitals. Further, we found various possibilities to grow and extend medical researcher's data sets in privacy-preserving ways, collaboration opportunities, technologies like Rmind that permit scalable, distributed data analysis, and others for automated data linking, de-identification and anonymization. However, we also identified issues where secure computation can not help, but that need to be solved on their own in the future in order to truly progress digitalization in health care. This includes missing or misaligned monetary incentives, lack of top-down enforcement of standardization, slow adoption of technologies due to lack of understanding of their necessity as well as their functionality and the optional, rare use of communication channels that result in health care institutions working in isolation. We thus deduce that the existing secure computation approaches in research and industry offer promising prospects but are not enough to cover clients' needs on their own.

Subsequently, we examine misalignments between secure computation offers and real-world health care and medical research requirements and reason what measures security companies and researchers could take to close these gaps. We determine that in research, usability is rarely taken into consideration and the companies need to take into account the skill set of their clients in order to foster marketability. Furthermore, public perception and security concerns play a significant role in the adoption of new technologies, which is why we propose to combine secure computation with existing compliance-based solutions to facilitate market entry, create more visibility and encourage transparency around these offers to reduce fear of contact. Moreover, cryptographers should consider the circumstances in which their protocols are used, for example in multi-party situations with many sub-parties that could cause issues even within otherwise trusted entities. Such considerations will hopefully increase robustness as well as performance and thus also advance the usability of the protocols. We conclude that solutions to the mentioned problems require a pragmatic attitude in research and client-centered product development.

Finally, we want to consider also the social responsibility of people inventing new technologies. Personal data should always be collected and processed responsibly and one needs to be aware that new, unprecedented kinds of (medical) research on vast amounts of data could have possibly unintended consequences. Simply applying cryptographic techniques does not make questionable research morally acceptable, and it is essential that ethics committees achieve a balance of appreciating additional privacy protections offered by cryptography without allowing it to be used to whitewash questionable proposals.

Appendix A

Appendix

A.1 Medical Files

As we have seen from comparing the data sources, there is a massive lack of standardization, the data is scattered and stored in silos. There are however, multiple competing standards on different levels, as there have been huge efforts in the past to drive towards interoperability. So far, none of them was a panacea to these issues, but it is crucial to understand their main selling points, what they are meant for and why some of them failed in order to talk about new approaches.

A.1.1 FAIR Guiding Principles

The FAIR Guiding Principles [123] are meant to be a guideline to support the reuse of research data. It states that to provide a good reusability, data has to be findable, accessible, interoperable and reusable. These principles are not only meant for humans managing data, but also for machine readability.

For each principle, the authors also state multiple subpoints that mainly focus on machine readability. For findability, these mainly consider the presence of metadata, indices and identifiers. Accessibility deals with clear communication protocols and authentication and authorization procedures. Interoperability manages the language, vocabulary and references used in the data collection. Finally, reusability means to handle data attributes, data usage licenses, standards and a clear origin of the data.

As we will see, existing data standards in health care frequently do not align well with these requirements.

A.1.2 Semantic Standards, Classification, Identifiers and Conventions

While many patient files still contain free text, graphs and images, we will here show some standards that have been at least partially adapted. We will focus on the few that during the research for this thesis seemed to be the most important in the Swiss hospital environment and only shortly mention the others we found, but the list probably still is not exhaustive.

The International Classification of Diseases (ICD) is a diagnostic tool maintained by the WHO. The current revision version is version 10, called ICD-10. Version 11 will be adopted in early 2022. The German modification of ICD-10, ICD-10 GM, is what is currently in use in Switzerland and has been translated into French and Italian for that matter. [106] ICD is a globally adopted standard for disease and health condition reporting. It promises easy sharing and comparing of reports and is suitable for disease monitoring, death cause, external causes of illness and more. [66] These properties might be limited in Switzerland because they are using their own modification together with Germany. While the ICD does not cover everything one might write in a patient record, it is a great and functional tool for the purpose it was designed. Its main drawback is the training of health care professionals to use it. Not only is it costly, but with each version, there are more terms that are even more specific, such that some physicians might not even know how to report unspecific statements from their patients. Even worse, the more complex a standard, the more likely the people using it will make mistakes. [14] Further, sometimes e.g. the cause of death of a patient with comorbidity is not certain. The physicians are then forced to decide what to put as the underlying cause of death and non-underlying cause of death, respectively. [65] The general lack of multi-dimensionality can lead to a distortion in the data set, makes databases hard to query (say you search for infections of the foot, you can only query infections or foot, and the results depend on in which hierarchy the attending doctor put it) and is also a reason why SNOMED CT has been introduced in Switzerland.

The Swiss Operation Classification (CHOP) is based initially on ICD-9 and is meant to classify treatment. [100] CHOP and ICD-10 are used by Swiss Health Care providers, especially hospitals, to get reimbursement from health insurances. [115] Thus, hospitals are automatically forced to use them. In ambulant care it is allowed to use International Classification of Primary Care (ICPC) or the Ticino Code, which are also for diagnosis and treatment classification, but much more coarse-grained and thus cost less time and training. [81] ICPC also allows to code why a patient visited a doctor, before they got a diagnosis. This is why general practitioners mostly use these two. There also exists conversion tables from ICPC-2 to ICD-10. [55] Because especially ICD-10 is so detailed and complicated, people are employed just to translate medical histories in ICD/CHOP. These do not have to be attending doctors but are responsible that the hospital gets its money.

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a machine readable set of medical terms. It codes diagnosis, clinical findings, surgical, therapeutic and diagnostic procedures, body structures, organisms, substances, pharmaceutical products, physical objects, physical forces, specimens and even the occupation of a patient. Its vast scope is meant to reduce the need for multiple systems in health records. Next to descriptions, it also offers concepts and relationships and thus builds a whole logical model. It was built for EHRs, with the goal to increase their efficiency. To be internationally operable, it was meant to be like a language, that in turn can be translated into other, human-readable languages. Mappings of existing standards to SNOMED CT allow the system to translate codings of different standards into SNOMED CT in the background, which in turn allows to export the codes into a third standard. [103, 13] Even though this system has been available in Switzerland since 2016, it is not adopted widely enough. The reason for this is a lack of incentives, since there is no law that enforces a better interoperability. [59] Same as for ICD-10, implementing it in an institution costs time and money. For the English version, SNOMED CT has NLP support which is increasingly extended. This means that free text can be entered and the clinical NLP extracts the necessary parts as SNOMED CT. [2] Being able to use free text in the reports would quite likely increase its adoption, since everyone is already used to it. In Switzerland, similar efforts are still in their early stages. [80] Since Germany joined SNOMED CT in January 2021 [48], there is hope that the now broader adoption will also drive efforts in Switzerland.

The Logical Observation Identifiers Names and Codes (LOINC) is probably one of the best-known coding systems, in Switzerland as well as internationally. Most EHRs are able to support it and Health Level 7 (HL7) made it part of its Clinical Document Architecture. In Switzerland, LOINC is mainly used to code laboratory results. Even though the country is in dire need of nationwide adoption because of its multilinguality, its own linguistic versions lack support because they do not cover enough terms. LOINC also has often been used for purposes it was not designed for, for example representing correlated data. Joining data that is LOINC coded remains challenging in many cases because of missing context and references. Backwards-compatibility is another challenge with LOINC, as the codes become increasingly detailed and specific, such that old, less specific codes would now map to multiple more specific ones. There are efforts to map LOINC with SNOMED CT to mitigate some of these issues. [30]

Other classification systems which are less important to us are presented in the following.

The Anatomical Therapeutic Chemical (ATC) Classification System is used in drug classification, mainly in pharmacology. It classifies ingredients, the system they act in, and their therapeutic, pharmacological and chemical properties. [90]

The Medical Dictionary for Regulatory Activities (MedDRA) was designed for registration, documentation and safety monitoring of medicinal products intended for human use. [122]

The North American Nursing Diagnosis Association (NANDA) is an association that designs a classification system for care diagnoses which is also called NANDA. It is coordinated with the Nursing Interventions Classification (NIC) and Nursing Outcomes Classification (NOC). [7]

A.1.3 Data Models

Next to the semantic standards, there is also a need for data models that describe how data should be structured, shared and especially accessed and queried. Multiple organizations work on standardized data models, some of which we will mention here to give the reader a better idea of where to look for further information.

The Clinical Data Interchange Standards Consortium (CDISC) developed a whole series of data sharing standards for data from clinical trials [23].

The Observational Medical Outcomes Partnership (OMOP) [97] offers a Common Data Model (CDM) for the systematic analysis of observational data. The idea is to have multiple databases with common format and representation, such that their data can easily be compared and analyzed.

Another standard is the Digital Imaging and Communications in Medicine (DICOM) [33] which is most commonly used to transmit medical images. It is the standard for medical imaging information and its related data and involves transmission, storage, retrieval, printing, processing and displaying medical images. Today, it has been widely adopted by hospitals and is increasingly also found in smaller practices.

An open-source approach is i2b2 [58], a clinical data warehousing and analytics research platform that enables sharing, integration, standardization and analysis. In its newer versions, it also makes use of the OMOP CDM.

The group that appears the most prominent is Health Level 7 (HL7) [54]. They provide a framework and standards for almost everything related to health information. All HL7 standards have ANSI/ISO/HITSP approvals. Two of its most popular standards are the Clinical Document Architecture (CDA) that provides a markup standard for clinical document structure and semantics, and the HL7 Fast Healthcare Interoperability Resources (FHIR) which facilitates the healthcare data exchange between all kinds of health care professionals. HL7 has collaboration agreements with the developers of SNOMED CT and LOINC, as well as several others.

An important pillar for most of these standards is the Resource Description Framework (RDF) [93], a very general approach to describing data and metadata. Everything is stored as a triple, consisting of subject, predicate and object. Here, the subject is a resource, the predicate describes its attributes and the relationship between subject and object. The goal of RDF is to represent interconnected data, which is something the classical semantics standards like ICD-10 lack.

The SPHN is also working on an RDF-based semantic interoperability framework, but because it is very new and still under development, we will only look at it later.

A.2 Legal Definitions

In this section, we present the most important definitions that frequently occur in legal frameworks, e.g. identifiability, anonymization or sensitive data. We will not only consider Swiss law but also GDPR and HIPAA.

The Federal Act on Data Protection (FADP) regulates the protection of privacy and the rights of a person whose data is processed by natural or legal persons. The Message on the Federal Act on Data Protection describes how the law should be interpreted, provides examples and historical context. We will also mention parts of the Swiss Human Research Act (HRA) and the Message on the Human Research Act. The HRA serves to protect human dignity and personality and complements certain parts of the FADP.

In the FADP, personal data is described as all information relating to an identified or identifiable person.

Sensitive personal data is personal data that is considered worthy of special protection and includes health data.

Next to identifiability, the FADP also mentions a personality profile, which is a ″collection of data that permits and assessment of essential characteristics of the personality of a natural person″. [110]

Identifiability is defined only in the Message on the Federal Act on Data Protection, as if a person is not clearly identified from the data alone, but can be inferred from the context or circumstances of the data.

However, this holds only if doable with manageable effort and does not enclose e.g. a resourceful statistical analysis. [111] Here, the interpretation of the law does also depends on the resources of the person having access to the data and whether they would, within reason, ″be able and willing to identify the subject″. [77]

The U.S. Health Insurance Portability and Accountability Act (HIPAA) offers much clearer mechanisms. [77] Their risk- and rule-based approaches will be discussed later.

The EU General Data Protection Regulation (GDPR) in turn reasons that to evaluate the identifiability of a person, all measures likely to be used (within reason) should be taken into account. This likelihood depends on factors such as the cost and timeliness, as well as available technologies. [77]

This is a slightly stronger statement than what we have seen from FADP, where we only dealt with ″manageable effort″, but the approach in general is similar, as it also does not state any more specific rules or technologies that should be considered in the process but binds the risk evaluation to its circumstances.

Note that GDPR affects institutions that offer their services to EU citizens, no matter where they are operating from. [92, 99]

GDPR further offers a definition for pseudonymization, which the FADP ambiguously calls ″verschlüsselt″, a term that can be translated with ″encrypted″ or ″coded″. We found that it means ″coded″ and refers to pseudonymization. [77]

GDPR defines pseudonymization as a modification of personal data such that the data cannot be linked to its data subject anymore without certain additional information that should be kept separately.

While FADP and GDPR remain silent about adequate measures, HIPAA presents two mechanisms for de-identification.
"Expert Determination" makes use of statistical methods that allow to modify data such that individuals cannot be identified anymore. The process consists of about three steps. First, the re-identification risk of the original data has to be evaluated. Then, applicable methods that reduce that risk have to be found and applied to the data. In the last step, the re-identification risk of the modified data set has to be evaluated. The new risk has to be "very small". The law relies on the expert to choose an appropriate interpretation of "very small", depending on the data set and its work environment.

"Safe Harbor" is a much simpler but also controversial approach. It presents a list of identifiers that have to be removed from an individual's file or changed. This list contains only 18 identifiers. The people modifying the data further should not have clear knowledge that the remaining data can be used for re-identification, e.g. the profession is none of the 18 identifiers, and some professions are quite unique, such that they too would have to be removed.

Even though both approaches state a certain awareness of a remaining re-identification risk, "Safe Harbor" has become controversial thanks to observations by Latanya Sweeney, who stated in 2000 that "87 percent of the U.S. population is uniquely identified by date of birth, gender, postal code" [108] and conducted further studies that heavily contest the privacy claims of "Safe Harbor". [109]

Bibliography

[1] *5 Top Multi-Party Computation (MPC) Solutions*. URL: https://www.startus-insights.com/innovators-guide/5-top-multi-party-computation-mpc-solutions/ (visited on 09/19/2021).

[2] *5.1 Natural Language Processing*. URL: https://confluence.ihtsdotools.org/display/docanlyt/5.1+natural+language+processing (visited on 09/19/2021).

[3] *Access to Data*. 2021. URL: https://clinicalstudydatarequest.com/Help/Help-Access-to-Data.aspx (visited on 09/19/2021).

[4] Rita Achermann, Pascal Düblin, and Ivan Nesic. *NLP-powered mapping of clinical reports onto SNOMED-CT concepts for tumour classification (NLPforTC)*. 2019. URL: https://sphn.ch/wp-content/uploads/2019/12/Abstract_RitaAchermann_May2019.pdf (visited on 09/19/2021).

[5] Amazon. *Security, Identity, and Compliance on AWS*. URL: https://aws.amazon.com/products/security/ (visited on 09/19/2021).

[6] Yongdae An et al. "Privacy-Oriented Technique for COVID-19 Contact Tracing (PROTECT) Using Homomorphic Encryption: Design and Development Study". In: *Journal of medical Internet research* 23.7 (July 2021), e26371–e26371. DOI: 10.2196/26371. URL: https://pubmed.ncbi.nlm.nih.gov/33999829.

[7] *Arbeitsblatt Pflegediagnosen*. 2020. URL: https://www.thieme.de/statics/dokumente/thieme/final/de/dokumente/tw_pflegepaedagogik/3-5-Pflegediagnosen.pdf (visited on 09/19/2021).

[8] Frederik Armknecht et al. *A Guide to Fully Homomorphic Encryption*. Cryptology ePrint Archive, Report 2015/1192. https://ia.cr/2015/1192. 2015.

[9] Arzneimittelkommission der deutschen Ärzteschaft (AkdÄ). *Nebenwirkungen melden: Ein Leitfaden für Ärzte*. 2019. URL: https://www.akdae.de/Arzneimitteltherapie/LF/PDF/Nebenwirkungen_melden.pdf (visited on 09/19/2021).

[10] Emmanuel Barraud. *Contact-tracing apps prove that they save lives*. 2021. URL: https://actu.epfl.ch/news/contact-tracing-apps-prove-that-they-save-lives/ (visited on 09/19/2021).

[11] Departement Klinische Forschung Universitätsspital Basel. *Patientenorientierte Forschung: Forschungskonsent – Information für Patientinnen und Patienten*. URL: https://www.unispital-basel.ch/en/teaching-research/forschung-mit-patientendaten/ (visited on 09/19/2021).

[12] *Become a clinical research volunteer*. URL: https://www.gsk.com/en-gb/research-and-development/trials-in-people/become-a-clinical-research-volunteer/ (visited on 09/19/2021).

[13] Heinz Bhenda and Christian Lovis. *Dürfen wir vorstellen? – SNOMED CT*. 2016. URL: https://saez.ch/journalfile/view/article/ezm_saez/fr/bms.2016.04896/da841f977f353329c3ab387ddb7bd3d6838636b3/bms_2016_04896.pdf/rsrc/jf (visited on 09/19/2021).

[14] Juerg Peter Bleuer and Hans Rudolf Straub. *Ausblick semantische Standards für eHealth in der Schweiz V 2.0*. Nov. 2015. URL: https://www.e-health-suisse.ch/fileadmin/user_upload/Dokumente/2016/D/160119_Mandatsbericht_Ausblick_semantische_Standards_Schweiz_D.pdf (visited on 09/19/2021).

[15] Dan Bogdanov, Sven Laur, and Jan Willemson. *Sharemind: a framework for fast privacy-preserving computations*. Full version of a paper that will be published at ESORICS 2008. db@math.ut.ee 14057 received 27 Jun 2008. 2008. URL: http://eprint.iacr.org/2008/289.

[16] Dan Bogdanov et al. "Rmind: A Tool for Cryptographically Secure Statistical Analysis". In: *IEEE Transactions on Dependable and Secure Computing* 15.3 (2018), pp. 481–495. DOI: 10.1109/TDSC.2016.2587623.

[17] Dan Bogdanov et al. *Students and Taxes: a Privacy-Preserving Study Using Secure Computation*. July 2016. DOI: 10.1515/popets-2016-0019.

[18] Peter Bogetoft et al. "Secure Multiparty Computation Goes Live". In: vol. 5628. Feb. 2009, pp. 325–343. ISBN: 978-3-642-03548-7. DOI: 10.1007/978-3-642-03549-4_20.

[19] Gaudet-Blavignac C et al. *A National, Semantic-Driven, Three-Pillar Strategy to Enable Health Data Secondary Usage Interoperability for Research Within the Swiss Personalized Health Network: Methodological Study*. 2021. URL: https://doi.org/10.2196/27591 (visited on 09/19/2021).

[20] Luca De Carli. *Schweiz: Ärzte melden Corona-Fälle per Fax*. Mar. 2020. URL: https://www.tagesanzeiger.ch/schweiz/standard/aerzte-muessen-coronafaelle-per-fax-melden/story/26352575 (visited on 09/19/2021).

[21] Sergiu Carpov et al. "Manticore: Efficient Framework for Scalable Secure Multiparty Computation Protocols". In: *IACR Cryptol. ePrint Arch.* 2021 (2021), p. 200.

[22] *CDC Interactive Database Systems*. URL: https://www.cdc.gov/surveillancepractice/data.html (visited on 09/19/2021).

[23] *CDISC*. URL: https://www.cdisc.org (visited on 09/19/2021).

[24] WHO Uppsala Monitoring Centre. *VigiLyze: Supporting national pharmacovigilance centres' quantitative and qualitative signal detection processes on national, regional and global levels*. URL: https://www.who-umc.org/vigibase/vigilyze/ (visited on 09/19/2021).

[25] Jung Hee Cheon et al. "Homomorphic Encryption for Arithmetic of Approximate Numbers". In: *Advances in Cryptology – ASIACRYPT 2017*. Ed. by Tsuyoshi Takagi and Thomas Peyrin. Cham: Springer International Publishing, 2017, pp. 409–437. ISBN: 978-3-319-70694-8.

[26] Moshinsky Chmiel. *A milestone for research in primary care in Switzerland: The FIRE project*. Jan. 2011. URL: https://smw.ch/article/doi/smw.2011.13142 (visited on 09/19/2021).

[27] MD Christian Lovis. *DeID (deidentification) of clinical narrative data in French, German and Italian and LOINC for Swiss Laboratories (L4CHLAB)*. 2019. URL: https://sphn.ch/wp-content/uploads/2019/12/Abstract_ChristianLovis_May2019.pdf (visited on 09/19/2021).

[28] *Clinical Study Data Request*. URL: https://clinicalstudydatarequest.com/Default.aspx (visited on 09/19/2021).

[29] The Federal Committee. *Bundesrat verabschiedet neue Zulassungskriterien für Leistungserbringer und vereinheitlicht Anforderungen an Spitalplanung*. 2021. URL: https://www.bag.admin.ch/bag/de/home/das-bag/aktuell/medienmitteilungen.msg-id-84111.html (visited on 09/19/2021).

[30] Fried-Michael Dahlweid, Matthias Kämpf, and Alexander Leichtle. "Interoperability of laboratory data in Switzerland: a spotlight on Bern:" in: *Journal of Laboratory Medicine* 42.6 (2018), pp. 251–258. DOI: doi:10.1515/labmed-2018-0072. URL: https://doi.org/10.1515/labmed-2018-0072.

[31] *Das FIRE-Projekt*. Apr. 2021. URL: https://www.hausarztmedizin.uzh.ch/de/fire2.html (visited on 09/19/2021).

[32] Statista Research Department. *Adoption of government endorsed COVID-19 contact tracing apps in selected countries as of July 2020*. 2021. URL: https://www.statista.com/statistics/1134669/share-populations-adopted-covid-contact-tracing-apps-countries/ (visited on 09/19/2021).

[33] *DICOM: Digital Imaging and Communications in Medicine*. URL: https://www.dicomstandard.org (visited on 09/19/2021).

[34] *Differential Privacy for Census Data Explained*. 2021. URL: https://www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx (visited on 09/19/2021).

[35] Susan Divald. *E-formalization case study: e-Estonia: A digital society for the transition to formality*. 2021. URL: http://www.ilo.org/wcmsp5/groups/public/---ed_emp/---emp_policy/documents/publication/wcms_781500.pdf (visited on 09/19/2021).

[36] Xiao Dong et al. "Developing High Performance Secure Multi-Party Computation Protocols in Healthcare: A Case Study of Patient Risk Stratification". In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2021 (May 2021), pp. 200–209. URL: https://pubmed.ncbi.nlm.nih.gov/34457134.

[37] *DP-3T Repository*. URL: https://github.com/DP-3T/documents (visited on 09/19/2021).

[38] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Found. Trends Theor. Comput. Sci.* 9.3–4 (Aug. 2014), pp. 211–407. ISSN: 1551-305X. DOI: 10.1561/0400000042. URL: https://doi.org/10.1561/0400000042.

[39] *e-estonia briefing centre*. URL: https://e-estonia.com/about-us/ (visited on 09/19/2021).

[40] T. Elgamal. "A public key cryptosystem and a signature scheme based on discrete logarithms". In: *IEEE Transactions on Information Theory* 31.4 (1985), pp. 469–472. DOI: 10.1109/TIT.1985.1057074.

[41] Danielle Enwood. *Zero-knowledge proofs – a powerful addition to blockchain*. June 2021. URL: https://blockheadtechnologies.com/zero-knowledge-proofs-a-powerful-addition-to-blockchain/ (visited on 09/19/2021).

[42] *Estonia creates a public code repository for e-governance solutions*. 2019. URL: https://e-estonia.com/code-repository-for-e-governance/ (visited on 09/19/2021).

[43] Kantonale Ethikkommission. *Daten, Proben & Datenschutz*. 2021. URL: https://www.zh.ch/de/gesundheit/ethik-humanforschung/daten-proben-datenschutz.html (visited on 09/19/2021).

[44] David Evans, Vladimir Kolesnikov, and Mike Rosulek. "A Pragmatic Introduction to Secure Multi-Party Computation". In: *Foundations and TrendsÂ® in Privacy and Security* 2.2-3 (2018), pp. 70–246. ISSN: 2474-1558. DOI: 10.1561/3300000019. URL: http://dx.doi.org/10.1561/3300000019.

[45] *FIRE: Bedeutung für die Forschung*. Apr. 2021. URL: https://www.hausarztmedizin.uzh.ch/de/fire2/bedeutungfuerdieforschung.html.

[46] Schweizerische Ethikkommission für die Forschung am Menschen. *Leitgedanken zu Registern in der Humanforschung*. 2019. URL: https://swissethics.ch/assets/pos_papiere_leitfaden/register_final_d.pdf (visited on 09/19/2021).

[47] *Gates Open Research Articles*. 2021. URL: ttps://gatesopenresearch.org/browse/articles (visited on 09/19/2021).

[48] *Germany's BfArM joins SNOMED International for nationwide use of SNOMED CT*. Jan. 2021. URL: https://www.snomed.org/news-and-events/articles/Germany-joins-SNOMEDCT (visited on 09/19/2021).

[49] Bundesamt für Gesundheit. *Strategie „eHealth" Schweiz*. 2007. URL: https://www.bag.admin.ch/bag/de/home/strategie-und-politik/nationale-gesundheitsstrategien/strategie-ehealth-schweiz.html (visited on 09/19/2021).

[50] Bundesamt für Gesundheit. *Technische Information: SwissCovid App: Einsatz von Bluetooth und den API von Apple und Google*. 2020. URL: https://www.bag.admin.ch/bag/de/home/krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/novel-cov/swisscovid-app-und-contact-tracing.html#-1591033202.

[51] O. Goldreich, S. Micali, and A. Wigderson. "How to Play ANY Mental Game". In: *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*. STOC '87. New York, New York, USA: Association for Computing Machinery, 1987, pp. 218–229. ISBN: 0897912217. DOI: 10.1145/28395.28420. URL: https://doi.org/10.1145/28395.28420.

[52] *Google Differential Privacy Repository*. URL: https://github.com/google/differential-privacy (visited on 09/19/2021).

[53] *Health Info Net AG*. URL: https://www.hin.ch (visited on 09/19/2021).

[54] *Health Level Seven International*. URL: https://www.hl7.org/about/index.cfm?ref=nav (visited on 09/19/2021).

[55] Dr. Walter Heckenthaler. *ICPC-2 – die Klassifizierung für die Primärversorgung*. Oct. 2017. URL: https://primaerversorgung.org/2017/10/19/icpc-2-die-klassifizierung-fuer-die-primaerversorgung/ (visited on 09/19/2021).

[56] *Hejasoftware*. URL: https://www.hejasoftware.ch (visited on 09/19/2021).

[57] *How to Securely Link Datasets from Different Organizations*. 2019. URL: https://go.privitar.com/2019-10-09-WW-1V-1140-how-to-securely-link-datasets-from-different-organizations-LandingPage.html (visited on 09/19/2021).

[58] *i2b2 Community Wiki*. URL: https://community.i2b2.org/wiki/ (visited on 09/19/2021).

[59] Sara Ibrahim. *Der weite Weg zur Digitalisierung von Patientendaten in der Schweiz*. Feb. 2021. URL: https://www.swissinfo.ch/ger/gesundheit_der-weite-weg-zur-digitalisierung-von-patientendaten-in-der-schweiz-/46373796 (visited on 09/19/2021).

[60] *IDASH PRIVACY & SECURITY WORKSHOP 2020*. 2020. URL: http://www.humangenomeprivacy.org/2020/ (visited on 09/19/2021).

[61] *IDASH PRIVACY & SECURITY WORKSHOP 2021*. 2021. URL: http://www.humangenomeprivacy.org/2021/about.html (visited on 09/19/2021).

[62] *Inpher*. URL: https://inpher.io (visited on 09/19/2021).

[63] *Inpher wins the iDASH Secure Genome Analysis Competition*. 2020. URL: https://inpher.io/news/inpher-wins-the-idash-secure-genome-analysis-competition/ (visited on 09/19/2021).

[64] *Institut Pasteur*. URL: https://www.pasteur.fr/en/public-health (visited on 09/19/2021).

[65] *International Classification of Diseases,Tenth Revision (ICD-10)*. Feb. 2020. URL: https://www.cdc.gov/nchs/icd/icd10.htm (visited on 09/19/2021).

[66] *International Statistical Classification of Diseases and Related Health Problems (ICD)*. 2021. URL: https://www.who.int/standards/classifications/classification-of-diseases (visited on 09/19/2021).

[67] Eric Blake Jackson. *e-Governance in Estonia: Balancing Citizen Data Privacy, Security and e-Service Accessibility*. 2021. URL: https://worldfinancialreview.com/e-governance-in-estonia-balancing-citizen-data-privacy-security-and-e-service-accessibility/ (visited on 09/19/2021).

[68] V. Junod. *Retrospective research: what are the ethical and legal requirements?* 2010. URL: https://doi.org/10.4414/smw.2010.13041 (visited on 09/19/2021).

[69] Laur Kanger and Pille Pruulmann-Vengerfeldt. *Usable and Efficient Secure Multiparty Computation: Deliverable D1.4 Expert Feedback on Prototype Application*. June 2014. URL: http://uaesmc.cyber.ee/files/D1.4-web.pdf.

[70] Lexie. *Zero-knowledge proofs explained: Part 1 and 2*. Feb. 2020. URL: https://www.expressvpn.com/blog/zero-knowledge-proofs-explained/.

[71] Poh Lian Lim. "Middle East respiratory syndrome (MERS) in Asia: lessons gleaned from the South Korean outbreak". In: *Transactions of The Royal Society of Tropical Medicine and Hygiene* 109.9 (Aug. 2015), pp. 541–542. ISSN: 0035-9203. DOI: 10.1093/trstmh/trv064. eprint: https://academic.oup.com/trstmh/article-pdf/109/9/541/5399428/trv064.pdf. URL: https://doi.org/10.1093/trstmh/trv064.

[72] Andrea Martania et al. *Data protection and biomedical research in Switzerland: setting the record straight*. Sept. 2020. URL: https://doi.org/10.4414/smw.2020.20332 (visited on 09/19/2021).

[73] *Meaningful Use: Electronic Health Record (EHR) incentive programs*. URL: https://www.ama-assn.org/practice-management/medicare-medicaid/meaningful-use-electronic-health-record-ehr-incentive (visited on 09/19/2021).

[74] *MedCo: Collective protection of medical data*. URL: https://medco.epfl.ch (visited on 09/19/2021).

[75] *Medication Errors*. URL: https://www.ema.europa.eu/en/human-regulatory/post-authorisation/pharmacovigilance/medication-errors (visited on 09/19/2021).

[76] Microsoft. *Microsoft Security Compliance Toolkit 1.0.* URL: https://docs.microsoft.com/en-us/windows/security/threat-protection/security-compliance-toolkit-10 (visited on 09/19/2021).

[77] Luca Dal Molin and Vera Rentsch. *Swiss Legal Framework for De-identification of Health-Related Data*. Dec. 2020. URL: https://sphn.ch/wp-content/uploads/2021/04/Homburger-memorandum_Swiss-Legal-Framework-for-De-identification-of-Health-Related-Data_20210105.pdf (visited on 09/19/2021).

[78] *MSF Research Protocols*. 2020. URL: https://fieldresearch.msf.org/handle/10144/241431 (visited on 09/19/2021).

[79] Damian Müller and Paul Rechsteiner. *Ein elektronisches Patientendossier für alle am Behandlungsprozess beteiligten Gesundheitsfachpersonen*. 2021. URL: https://www.parlament.ch/centers/kb/Documents/2019/Kommissionsbericht_SGK-S_19.3955_2021-02-22.pdf.

[80] *NLP-powered mapping of clinical reports onto SNOMED-CT concepts for tumour classification (NLPforTC)*. 2019. URL: https://sphn.ch/seminar-training/nlp-powered-mapping-of-clinical-reports-onto-snomed-ct-concepts-for-tumour-classification-nlpfortc/ (visited on 09/19/2021).

[81] P.Meier. *Wieviel muss oder darf die Krankenkasse wissen?* 2004. URL: https://saez.ch/journalfile/view/article/ezm_saez/de/saez.2004.10501/80321b82621ccaedcd1ba598d280d1cad13f357d/saez_2004_10501.pdf/rsrc/jf (visited on 09/19/2021).

[82] Danioreac Paola et al. *The SwissCovid Digital Proximity Tracing App after one year: Were expectations fulfilled?* 2021. URL: https://doi.org/10.4414/SMW.2021.w30031.

[83] Caroline Criado Perez. *Invisible Women: Data Bias in A World Designed for Men*. 2019.

[84] Dr. med. Peter Kleist. *Kantonale Ethikkommission*. 2021. URL: https://www.zh.ch/de/gesundheitsdirektion/ethikkommission.html (visited on 09/19/2021).

[85] Federico Plantera. *The cornerstone of e-governance is trust*. 2018. URL: https://e-estonia.com/cornerstone-governance-trust/ (visited on 09/19/2021).

[86] Praetorian. *The Elephant in the Room: Why Security Programs Fail*. 2021. URL: https://lp.praetorian.com/why-security-programs-fail (visited on 09/19/2021).

[87] *Privitar*. URL: https://www.privitar.com (visited on 09/19/2021).

[88] Privitar. *Managing the Lifecycle of Sensitive Data with the Privitar Data Privacy Platform*. URL: https://alphazetta.ai/wp-content/uploads/2021/03/Managing-the-Lifecycle-of-Sensitive-Data-Platform-Whitepaper-A4.pdf.

[89] *Project Mapping*. 2019. URL: https://sphn.ch/network/project-overview/ (visited on 09/19/2021).

[90] *Purpose of the ATC/DDD system*. 2018. URL: https://www.whocc.no/atc_ddd_methodology/purpose_of_the_atc_ddd_system/ (visited on 09/19/2021).

[91] *RARE DISEASES DEVELOPMENT PLAN*. 2014. URL: https://www.sm.ee/sites/default/files/content-editors/eesmargid_ja_tegevused/Tervis/Tervishoiususteem/harvikhaiguste_arengukava_en.pdf (visited on 09/19/2021).

[92] *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/EC (General Data Protection Regulation) (Text with EEA relevance)*. May 2016. URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32016R0679.

[93] *Resource Description Framework (RDF) Model and Syntax Specification*. 1999. URL: https://www.w3.org/TR/PR-rdf-syntax/Overview.html (visited on 09/19/2021).

[94] R. L. Rivest, A. Shamir, and L. Adleman. "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems". In: *Commun. ACM* 21.2 (Feb. 1978), pp. 120–126. ISSN: 0001-0782. DOI: 10.1145/359340.359342. URL: https://doi.org/10.1145/359340.359342.

[95] Dr. Ramon Saccilotto. *E-General Consent: Development and Implementation of a Nationwide Harmonized Interactive Electronic General Consent*. 2019. URL: https://sphn.ch/wp-content/uploads/2019/12/Abstract_RamonSaccilotto_March2019.pdf (visited on 09/19/2021).

[96] P. Samarati and L. Sweeney. *Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression*. Tech. rep. 1998. URL: http://www.csl.sri.com/papers/sritr-98-04/.

[97] Observational Health Data Sciences and Informatics. *OMOP Common Data Model*. URL: https://www.ohdsi.org/data-standardization/the-common-data-model/ (visited on 09/19/2021).

[98] *Secure Computing Platform*. URL: https://sharemind.cyber.ee/secure-computing-platform/ (visited on 09/19/2021).

[99] EFTA Sekretariat. *General Data Protection Regulation (GDPR) entered into force in the EEA*. 2018. URL: https://www.efta.int/EEA/news/General-Data-Protection-Regulation-GDPR-entered-force-EEA-509576 (visited on 09/19/2021).

[100] Bundesamt für Statistik (BFS) Sektion Gesundheit der Bevölkerung Bereich Medizinische Klassifikationen. *Schweizerische Operationsklassifikation (CHOP), Systematisches Verzeichnis – Version 2021*. 2020.

[101] Adi Shamir. "How to Share a Secret". In: *Commun. ACM* 22.11 (Nov. 1979), pp. 612–613. ISSN: 0001-0782. DOI: 10.1145/359168.359176. URL: https://doi.org/10.1145/359168.359176.

[102] *Sharemind: The next generation of data-driven services with end-to-end data protection and accountability*. URL: https://sharemind.cyber.ee (visited on 09/19/2021).

[103] *SNOMED CT 5 step briefing*. 2021. URL: https://www.snomed.org/snomed-ct/five-step-briefing (visited on 09/19/2021).

[104] *SPHN WHERE DO WE STAND TODAY*. 2020. URL: https://sphn.ch/wp-content/uploads/2020/11/201112_SPHN_Factsheet_web_DEF.pdf (visited on 09/19/2021).

[105] Radames Cruz Moreno Sreekanth Kannepalli Kim Laine. *Password Monitor: Safeguarding passwords in Microsoft Edge*. Jan. 2021. URL: https://www.microsoft.com/en-us/research/blog/password-monitor-safeguarding-passwords-in-microsoft-edge/ (visited on 09/19/2021).

[106] Bundesamt für Statistik. *Medizinische Kodierung und Klassifikationen*. URL: https://www.bfs.admin.ch/bfs/de/home/statistiken/gesundheit/nomenklaturen/medkk.html (visited on 09/19/2021).

[107] *Strategie eHealth Schweiz 2.0*. 2018. URL: https://www.bag.admin.ch/bag/de/home/strategie-und-politik/nationale-gesundheitsstrategien/strategie-ehealth-schweiz.html (visited on 09/19/2021).

[108] Latanya Sweeney. *Simple Demographics Often Identify People Uniquely (Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000)*. 2000. URL: https://dataprivacylab.org/projects/identifiability/paper1.pdf (visited on 09/19/2021).

[109] Latanya Sweeney et al. “Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study.” In: *Technol Sci* 2017 (2017).

[110] The Federal Council of the Swiss Confederation. *Federal Act on Data Protection*. Mar. 2019. URL: https://www.fedlex.admin.ch/eli/cc/1993/1945_1945_1945/en.

[111] The Federal Council of the Swiss Confederation. *Message on the Federal Act on Data Protection*. Mar. 1988.

[112] The Federal Council of the Swiss Confederation. *Message on the Federal Act on Research Involving Human Subjects*. Oct. 2009.

[113] *Swiss Personalized Health Network (SPHN): Infrastructure building to enable nationwide use and exchange of health data for research*. URL: https://sphn.ch (visited on 09/19/2021).

[114] *SwissCovid-App-Monitoring*. URL: https://www.experimental.bfs.admin.ch/expstat/de/home/innovative-methoden/swisscovid-app-monitoring.assetdetail.13407769.html (visited on 09/19/2021).

[115] SwissDRG. *Fixed rate per case payments in Swiss hospitals: Basic information for healthcare professionals*. 2015. URL: https://www.swissdrg.org/application/files/1815/0234/7188/170810_SwissDRG_Brochuere_e.pdf (visited on 09/19/2021).

[116] Swissmedic. 2021. URL: https://www.swissmedic.ch/swissmedic/de/home/humanarzneimittel/marktueberwachung/pharmacovigilance.html (visited on 09/19/2021).

[117] *swissmedic: Schweizerisches Heilmittelinstitut*. URL: https://www.swissmedic.ch/swissmedic/de/home.html (visited on 09/19/2021).

[118] Dr. Matthias Templ. *Anonymisation of data sets from Helsana*. URL: https://www.zhaw.ch/en/research/research-database/project-detailview/projektid/2453/ (visited on 09/19/2021).

[119] *Tumorboard*. URL: https://www.usz.ch/zuweisende/tumorboards/ (visited on 09/19/2021).

[120] *Unlocking siloed data for the NHS*. 2020. URL: https://go.privitar.com/rs/588-MYA-374/images/EN_CS_2020_NHS.pdf (visited on 09/19/2021).

[121] My Villius Zetterholm, Yanqing Lin, and PÃ€ivi Jokela. "Digital Contact Tracing Applications during COVID-19: A Scoping Review about Public Acceptance". In: *Informatics* 8.3 (2021). ISSN: 2227-9709. DOI: 10.3390/informatics8030048. URL: https://www.mdpi.com/2227-9709/8/3/48.

[122] *Vision for MedDRA*. URL: https://www.meddra.org/about-meddra/vision (visited on 09/19/2021).

[123] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (2016), p. 160018. DOI: 10.1038/sdata.2016.18. URL: https://doi.org/10.1038/sdata.2016.18.

[124] *Wyden, Rubio, Warner Introduce "Student Right to Know Before You Go Act" to Empower Students as Consumers and Showcase New Privacy-Protecting Technology*. 2017. URL: https://www.wyden.senate.gov/news/press-releases/wyden-rubio-warner-introduce-student-right-to-know-before-you-go-act-to-empower-students-as-consumers-and-showcase-new-privacy-protecting-technology (visited on 09/19/2021).

[125] Andrew C. Yao. "Protocols for secure computations". In: *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*. 1982, pp. 160–164. DOI: 10.1109/SFCS.1982.38.

[126] Mark Zastrow. *South Korea is reporting intimate details of COVID-19 cases: has it helped?* 2020. URL: https://www.nature.com/articles/d41586-020-00740-y (visited on 09/19/2021).

[127] Eli Zimmerman. *How Multiparty Computation Could Secure Higher Ed Data Sharing*. 2018. URL: https://edtechmagazine.com/higher/article/2018/07/how-multiparty-computation-could-secure-higher-ed-data-sharing (visited on 09/19/2021).