Discovering health disparities: Designing a secure multiparty architecture for social health research

Brontë Kolar
Responsible Professor: Zekeriya Erkin
Cyber Security Group, Department of Intelligent Systems, Delft University of Technology

Abstract

Social determinants such as a person's race, level of education, and income can be responsible for their health outcomes. Consequently, we see that discrimination along the social spectrum results in health disparities. In an effort to close the gaps in healthcare systems, these determinants have been heavily researched. Open questions remain regarding their underlying mechanisms, which can potentially be answered by combining government data from social sectors with healthcare data. Under modern data legislation, such as the General Data Protection Regulation in the European Union, it is challenging to combine these government datasets for such research purposes. Multiparty computation (MPC), a cryptographic technique that allows for two or more parties to securely compute a function over data, opens the door for siloed government datasets to be combined and analyzed in a manner compliant with data legislation. This paper presents survey data from experts that supports the feasibility of using MPC to securely investigate the social determinants of health, as well as a potential architecture based on additive secret sharing that could be utilized by governments to investigate these determinants. This is the first formal research into this application of MPC and it aims to evaluate how new developments in cryptography can be leveraged to advance health equity and bring justice to those systemically discriminated against.

1 Introduction

Amidst clinical advancements in medical care, it can be forgotten that social factors continue to contribute to 20% of the health outcomes of individuals [1]. Research has shown that factors including race, level of education, and income affect the way people live and their consequent chance of illness [2]. As ethnic minorities are over-represented in COVID-19 cases and deaths, the recent pandemic stands as a testimony of these inequalities. In Chicago alone, over 50% of COVID-19 cases and almost 70% of COVID-19 fatalities are disproportionately within the black population, who only make up 30% of the city's population [3]. Discrepancies along the social spectrum affect the distribution of healthcare and health equity cannot be achieved until policies are derived that account for these structural inequities [4].

The non-medical factors that shape health outcomes are known commonly as the social determinants of health. Despite progress in documenting and understanding these social determinants, open questions regarding the mechanisms underlying their impact on health remain numerous [5]. With further insight into the social determinants of health, policies and programs can be enacted that influence these determinants to improve health equity [6]. A barrier to understanding these factors is the difficulty in obtaining data across sectors, such as education, housing, and labour. Analyzing social data in combination with healthcare data could improve understanding in this area and ultimately allow for impactful intervention [5].

Analysing cross-sectoral data is prevented by data policies, many of which are outdated and do not take into account the security capabilities of modern data sharing methods [5]. Due to the sensitive nature of health data, it is often highly protected by governments, who further place limits on the data's ability to be shared. This presents a crossroad; either relaxing legislature to allow for a broader scope of legal data sharing at the risk of an increase of security breaches, or implementing secure frameworks to allow for privacy-preserving sharing. There has been a movement across countries to have greater data sharing between siloed ministries, such as the Ontario Public Service Data Integration Framework in Canada, a new inter-ministry data sharing project that focuses on a legislative approach to greater collaboration [7]. Simultaneously, recent advances in cryptography have brought about new solutions that allow for secure collaboration to investigate the social determinants of health.

Despite general industry unawareness, multiparty computation (MPC) presents an opportunity to share datasets securely. Since the first works of MPC were published, improvements in efficiency make it now possible to use these protocols in industry [8]. With MPC, national datasets can be analyzed without ever revealing sensitive information about said data, allowing for privacy-preserving statistics that could be leveraged to shine light on the social determinants of health. By leveraging MPC, data can be collected and processed in a secure manner that is compliant with the majority

of current global data legislation, such as the General Data Protection Regulation (GDPR) in the European Union [9].

Health research institutions are also beginning to investigate methods of privacy-preserving record linkage (PPRL) with MPC to share and analyze healthcare data. In 2018, a research group utilized the Sharemind platform to use MPC to remove duplicate records from multiple medical databases without violating the privacy of the contributing data centers or the patient information stored in the records [10]. While examples of MPC in industry vary in their security guarantees, implementations, and communication models, solutions have been deployed in domains ranging from analyzing gender pay gaps [11] to genomic association analysis [12], supporting the potential for MPC in industry.

This paper assesses how MPC can be leveraged to unite government datasets to perform privacy-preserving record linkage and secure statistical analysis, allowing for large-scale analysis of the social determinants of health. After outlining background information on relevant topics such as data legislature and MPC techniques in Section 2, related works in this field are discussed in Section 3. Section 4 states the chosen methods of research. Through discussions with experts and healthcare institutions, the feasibility of MPC being used in this domain is presented in Section 5 through statistics gathered through a primary research survey. Based on requirement elicitation from these survey results, an architecture that utilizes MPC to efficiently investigate the social determinants of health is outlined in Section 6, along with its limitations and broader considerations in Section 7.

2 Building Blocks

Secure social health research is a complex issue rooted in social justice and restricted by outdated data legislation. Modern cryptographic techniques have the potential to push the boundaries of research and support in combining government datasets to investigate the social determinants of health. This section outlines relevant background information regarding the current techniques for linking and analysing records in industry, and common multiparty computation techniques.

2.1 Privacy-Preserving Record Linkage

Multiple organizations, such as the Institute for Clinical Evaluative Sciences (ICES) in Canada, are experimenting on conducting secure health research using a process known as privacy-preserving record linkage (PPRL). A key step involved in combining datasets involves linking records and this technology aids in multiple parties sharing related records. PPRL addresses problems in data sharing by linking records that correspond to the same entity across different datasets that are to be combined [13]. One main challenge of using PPRL for Big Data applications is preserving the privacy and confidentiality of the entities in the submitted databases. Modern PPRL techniques often represent a compromise between privacy and scalability, as scalability is increasingly important and taken into account for larger applications [14].

Current PPRL methods include secure hash encoding, statistical linkage keys, and bloom filters [15] [16] [17]. Multiparty computation (MPC) protocols have also been used for PPRL but are generally less efficient. Despite these computational limitations, they offer clear security guarantees over other options, such as data perturbation methods or Bloom filters [18] [19]. Common MPC techniques for PPRL include secure set union, secure set intersection, and secure scalar product [16].

2.2 Multiparty Computation

Multiparty computation (MPC) is a cryptographic technique that enables the privacy-preserving computation of a function. Two or more parties collectively compute a function on data, such that each party learns nothing about the private inputs of the other parties. In essence, the data input of one party is hidden from all other parties. Optionally, these input parties can also receive the output [20].

Some MPC protocols propose a client-server model in which data shares are sent by clients to servers that jointly compute a function. Compared to peer-to-peer models, client-server instances have lower network communication requirements and allow for reduced overhead [21]. Data is first encrypted by randomly splitting data into secret shares. On its own, each share reveals no information about the data. These shares are then distributed over servers controlled by different trustees, which then perform the necessary computation jointly. The privacy of the system is ensured as long as a subset of the trustees acts honestly and the data is fully protected both at-rest and at-use [8] [22].

MPC consists of the following three separate roles [23]:

1. **Input Parties:** Input parties provide the datasets in which computation will be done over.
2. **Computation Parties:** Computation parties are responsible for executing privacy-preserving computation on the data provided by input parties on servers. These computation parties are not able to construct the individual inputs and solely deploy the computation securely.
3. **Result Parties:** Finally, there are result parties who learn the results of the computation.

These parties can intersect and different stakeholders may take on different roles, such as an input party also receiving the results or serving as one of the computation parties [24]. It has been demonstrated that three computation parties produce the fastest MPC protocols [10]. Common MPC protocols include additive secret sharing and garbled circuits.

2.3 Additive Secret Sharing

Secret sharing schemes are used to securely distribute private values to a group of parties. They involve dividing a secret value into multiple shares and distributing these shares across multiple parties, such that a single party is prevented from having complete knowledge of the secret [25]. Additive secret sharing involves breaking a numeric secret into fragments that add up to the original secret. To reconstruct the secret, all parties must combine their shares together to reveal the original secret value.

This concept is based on additive homomorphic encryption, a scheme in which if $A = E(x)$ and $B = E(y)$, then $A \bigotimes B = E(x + y \mod p)$, where p is some integer modulus and $\bigotimes$ is a well-defined operation [26]. An additive secret

sharing protocol allows a secret m to be shared among n parties. These algorithms choose n strings $(s_1, ..., s_n)$ uniformly at random subject to the requirement that

$$\sum_{i=1}^{n} s_i = m \mod p. \quad (1)$$

An example of additive secret sharing is the Sharemind framework, which uses additive sharing over $\mathbb{Z}_{2^{32}}$ where a secret value s is split to shares $s_1, ..., s_n \in \mathbb{Z}_{2^{32}}$ such that $s_1 + ... + s_n$ is identical to $s \mod 2^{32}$ and any $n-1$ element subset $s_{i_1}, ..., s_{i_{n-1}}$ is uniformly distributed. Sharemind uses a 32-bit architecture to achieve maximal efficiency because integer arithmetic in most modern computers is based on a 32-bit architecture. It is most time and space efficient to use additive secret sharing schemes over $\mathbb{Z}_{2^{32}}$. However, there exists a tradeoff, as secret sharing schemes based on Shamir secret sharing fail over 2^{32}. To tackle this, Sharemind implemented their own multiplication protocol that converts between $\mathbb{Z}_2$ and $\mathbb{Z}_{2^{32}}$. Inputs are gathered as shares over $\mathbb{Z}_2$ to avoid fraudulent inputs and these shares are later converted to get corresponding shares over $\mathbb{Z}_{2^{32}}$. Through this approach, participants cannot learn anything about s unless all of them join their shares [25]. Additive secret sharing is a common approach for an honest majority of participants [8].

Consider an architecture with three computing parties that conduct computations over three servers. The private representation of a record r can be defined as $r = (r_1, r_2, r_3)$, where r_i denotes a secret share held by the i-th server. In an additive secret sharing scheme, there is a constraint that these shares sum up to r. Thus, if anyone has one or two shares of r, they cannot infer anything about r [10]. Protocols that support basic arithmetic, relation operations, logical operations, addition, multiplication, comparisons, and more have been created that transform shares of the input into shares of the output [27].

2.4 Garbled Circuits

Nearly every MPC protocol that allows one to evaluate a Boolean circuit is based on Yao's garbled circuits [26]. Garbled circuits enable two-party secure computation in which two parties can jointly evaluate a function over their private inputs. The function must be described as a Boolean circuit in a garbled circuit protocol [28]. While garbled circuit-based MPC is generally less computationally efficient than secret sharing-based MPC, it allows for secure computation with a constant number of rounds, which is especially relevant for protocols executing over the Internet. However, it is limited to simple Boolean operations and there is work to do in this sub-field with respect to increasing efficiency [12].

2.5 Legal Limitations Investigating the Social Determinants of Health

In many countries, sharing health data across several organizations is limited by laws or regulations, such as the GDPR in the European Union or the Personal Health Information Protection Act (PHIPA) in Ontario, Canada [9] [29]. Due to the sensitivity of health data, there are often further regulations regarding sharing it. For example, under PHIPA, it is extremely difficult to share government health datasets with other parties, such as government revenue agencies, for large-scale research purposes. However, techniques like MPC comply with many strict data regulations and open the door for data to be leveraged. Data remains encrypted at-rest and at-use, which is a paradigm shift from the scope of current data standards. MPC is compliant with the GDPR, as the output of the initial encryption step of MPC, where the data is divided into shares, is considered non-personal data due to MPC's unique data protection properties [22].

3 Related Works

A number of relevant research endeavors have been completed that support the feasibility of using MPC to combine and analyze datasets while preserving their security. A study conducted by the Estonian government in 2016 supports the potential of using MPC for large-scale statistical studies on protected government data. The study linked a tax payment database from the Estonian Tax and Customs Board and a higher education database from the Ministry of Education and Research. Data collection and analysis were conducted using Sharemind, a secure multiparty computation system that provides cryptographic protection. The deployment diagram of Sharemind from [25] can be viewed in Figure 1. While the scalability of MPC is viewed as a limitation, this study successfully utilized ten million tax records and half a million education records in the analysis, demonstrating that it is possible to run a private statistical study on large sets of real government data [24].

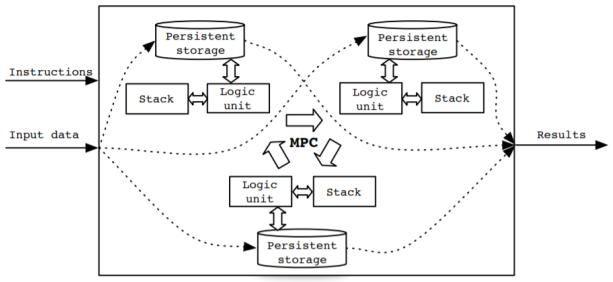


Figure 1: Deployment diagram of Sharemind platform, which is based on additive secret sharing. The input data and computation instructions are delivered to servers that use multi-party computation techniques to execute programs step by step and finally return the end results [24].

Further work has been conducted with regards to privacy-preserving record linkage of large databases using MPC. A Sharemind implementation was used to combine data from 1000 medical centers with 10 000 records each, giving accurate results and computation times that would be acceptable in real-world scenarios [10]. Further works also support that these protocols can be further optimized by MPC-optimized block ciphers and aggregated operations [30].

While the Sharemind platform leverages additive secret sharing, platforms that support secure collaborative analyt-

ics of SQL queries with strong security guarantees using garbled circuits have been developed. Senate is a new garbled circuit-based platform that prevents data leaks, even if $m-1$ of m parties fully collude. This new platform leverages new insights in query planning and cryptographic techniques to achieve efficient results. The Senate researchers developed a new protocol called secure MPC decomposition, which includes securely decomposing a large cryptographic computation into smaller and parallel computations, planning an efficient decomposition, and assigning a part of the query to local computation. While not suited for all queries, Senate can improve query runtime by 145 times for certain applications [31].

Achieving health equity is a topic of growing concern and governments are increasingly looking to fill gaps in their healthcare systems. Cryptographic technologies have potential to shape social issues and MPC has been leveraged in the past to expose social inequalities. In 2017, a web application for secure data analysis focused on usability was used by the Boston Women's Workforce Council to prove that the gender pay gap in Boston was much larger than previously estimated. Companies in Boston were able to securely contribute their data to this study and were able to take part without having access to the input data of other parties or the results of the analysis. Companies and institutions generally do not want to provide their data to be used for such studies, as they could incur legal liabilities, but MPC has opened the door for information to be collected securely and never be traced back to the institutions who provided it [11].

To the best of our knowledge, this paper presents the first formal research on the feasibility of using MPC to investigate the social determinants of health. While this is not the first case of an MPC implementation being used on government data, it is the first paper that looks at MPC's relevance in social health research.

4 Method

In order to gauge the perceived need and feasibility of MPC in addressing the search for social determinants of health, interviews were conducted with domain experts, following an extensive literature review. In parallel, a survey was used to gain quantitative data on the applicability of MPC for researching the social determinants of health.

4.1 Literature Review

A literature review was conducted to evaluate current methods of investigating social determinants of health and the potential of MPC protocols to support social health research. This also involved collaborating and sharing relevant papers with other researchers investigating the relevance and potential of MPC in various industry domains.

4.2 Expert Interviews and Survey

Following this literature review, interviews were conducted with experts in order to gain information on the limitations and current state-of-the-art techniques for researching these social determinants. This included experts in healthcare, policy and data security from organizations including Ontario Health, the Institute for Clinical Evaluative Sciences (ICES), University of Maastricht Research Hospital, and Meldpunt Discriminatie Regio Amsterdam. Through these interviews, insight into this use-case was gained, as well as industry challenges and standards that allowed for requirement elicitation. Finally, a survey was created with Google Forms and shared with these stakeholders to record quantitative data on the feasibility of MPC.

Ultimately, the knowledge gained through this method was used to design a secure and policy-compliant architecture that allows for investigating the social determinants of health.

5 Survey Results

5.1 Response Overview

Ten responses were collected from professionals across the fields of healthcare and data privacy, including researchers and senior managers at healthcare and research organizations. The full list of survey questions and their corresponding responses can be found respectively in sections A.1 and A.2 of the Appendix.

Participants work across Canada, the United Kingdom, and the Netherlands. Eight of these participants were familiar with the social determinants of health (80%) and seven were familiar with MPC (70%). The breakdown of the industry participants work in and their familiarity with relevant research concepts is depicted visually in Figure 2.

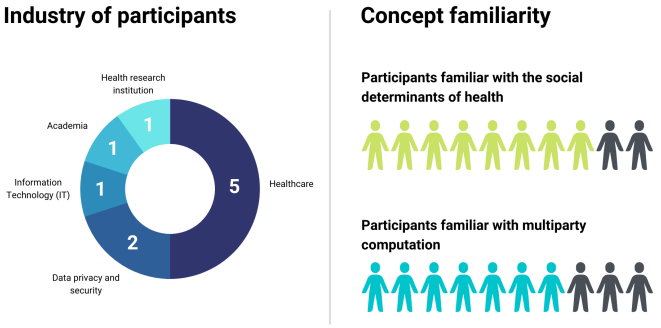


Figure 2: Industry breakdown of participants and their familiarity with the social determinants of health and MPC.

Regarding the social determinants of health, main findings include:

- Nine participants believe governments should do more to investigate social determinants of health.
- All participants stated that they believe if government organizations shared more data between each other, larger investigations into social determinants of healthcare could be conducted.
- All participants stated that information on the social determinants of health is important for healthcare and government systems.
- The main stated barriers to furthering research in this domain were reported to be legislature (50%), willingness

to investigate (20%), funding (10%), and political motivation (10%). Technology was not reported as a barrier.

Regarding MPC being leveraged to investigate these determinants, main findings include:

- All participants thought MPC could be used to combine data across institutions to investigate the social determinants of health.
- All participants were neutral or believed using MPC to combine data across institutions to investigate the social determinants of health is a feasible solution (1-5 scale, mean = 3.9). See Figure 3.
- All participants were neutral or believed MPC is a better alternative solution to current methods of investigating these determinants (1-5 scale, mean = 4.1). See Figure 4.
- The reported barriers reported by participants that stand in the way of adopting this technology include legislative/regulatory delays (100%), distrust in the technology (50%), lack of financial support (50%), explainability (10%), awareness (10%), lack of user tech-savviness (10%), concern it is too new (10%).
- The reported reasons why organizations are not already using MPC include lack of government motivation (60%), lack of awareness (50%), lack of funding (50%), and the technology not being perceived to be necessary (50%).

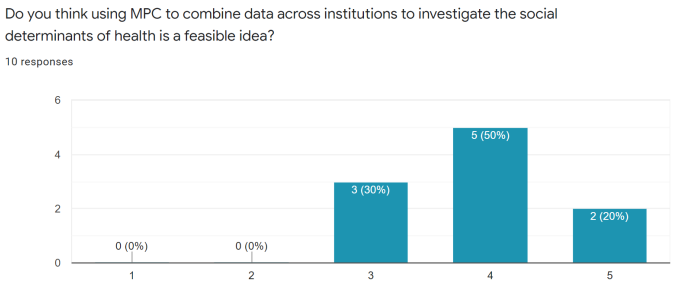


Figure 3: Survey results regarding the feasibility of using MPC to investigate the social determinants of health (*1 = not feasible, 5 = very feasible*).

5.2 Further Survey Findings

Participants recognised the benefits of building a new MPC architecture to investigate the social determinants of health. One participant reported that leveraging MPC would make it easier to combine sensitive data from different sources. Despite this perceived advantage, other participants stated that working with other organizations that would have the responsibility to implement and support an MPC solution in their data environment brings concerns about this technology being implemented in a timely manner. A third participant stated that an MPC solution offered as a third-party service with all required regulatory controls in place would significantly speed up the process of adoption.

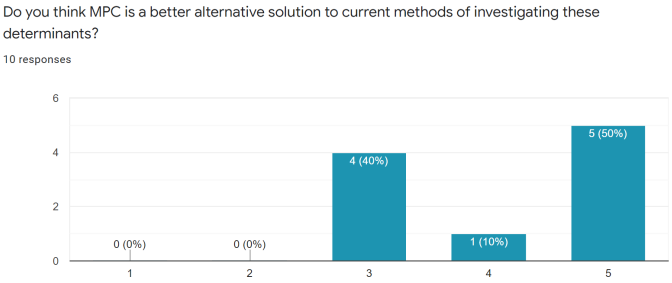


Figure 4: Survey results regarding if MPC is a better alternative to current methods of investigating the social determinants of health (*1 = worse than current methods, 5 = better than current methods*).

Regarding the feasibility of MPC in this domain, a participant reported that their experience with using homomorphic encryption for PPRL with healthcare data involved scalability concerns that made it infeasible for their organization.

5.3 Elicited Requirements

From concerns stated in this survey, three important values can be elicited for this architecture - *computation efficiency*, *security* and *ease of adoption*. Stakeholders are proceeding with legislative changes to ease intergovernmental collaboration, rather than technological. Partly due to lack of awareness, this can also be driven by the perceived challenge of developing and implementing this technology. Ensuring accessibility, efficiency, and security concerns are addressed through an MPC solution seem to be the most relevant architecture requirements.

To allow for computational efficiency, a potential architecture must minimize overhead communication to reduce total computation time. This involves incorporating state-of-the-art optimizations into current additive sharing approaches or even investigating potential cases for garbled circuits. An MPC architecture that accepts queries in a common language, such as SQL could ease the adoption process and decrease the time it takes for new data analysts to convert to a new system.

6 Architecture Solution

6.1 Architecture Overview

The following solution outlines an MPC architecture based on 3-party additive secret sharing. The architecture accepts SQL queries and supports multiple parties submitting private inputs that are to be computed by 3 computing parties. Aggregation operations are parallelized and MPC-optimized block ciphers are used in order to increase efficiency. This architecture is depicted in Figure 5, where the stages of data sharing, computation, and results can be viewed.

It has been observed that the architecture for which the fastest MPC protocols are known and mature platforms exist is for three computation parties [10]. This solution assumes that there is an honest majority, where a maximum of 1 of 3 servers can be corrupted. It is also assumed that the adversaries are passive and the corrupted server follows the

outlined protocol, despite accessing all messages received by the server.

Collaborating institutions for investigating the social determinants of health are housed in different government sectors, so a semi-honest model is acceptable. Government parties likely do not have competing interests and would not benefit from being corrupt, as is the case sometimes with private sector collaboration.

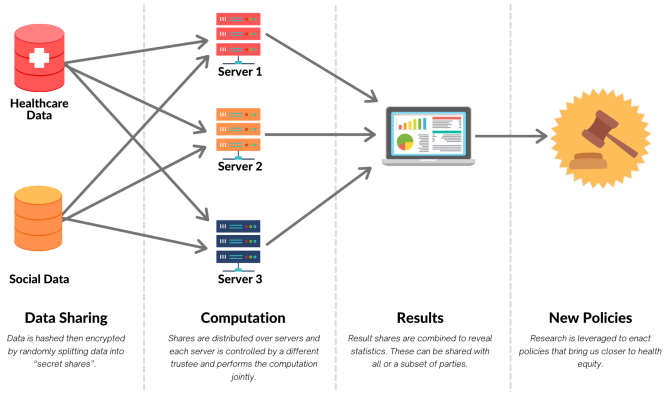


Figure 5: Proposed MPC solution architecture based on additive sharing.

6.2 Stakeholders and Deployment

Input parties are composed of various stakeholders that wish to share data to investigate the social determinants of health. Common cases include governmental health datasets being combined with datasets from another non-health ministry, such as those from a tax authority, education ministry or transportation ministry. While the architecture proposed has two input parties, more can be added if analyses are to be run over the data of more than two sectors. This would still require three computing servers.

Two input parties will also serve as a computing party. For the third computing party, an external third government server could be used to maintain trust and would be provided by another authority. Result parties can include input parties, as well as other organizations that want to receive the outputted data.

6.3 Data Pre-processing

Data owners collect information and store it in a way that supports their internal analyses and operations. Often, this means that datasets are not stored in a way that makes them analyzable with other data sets. For example, name data or dates of birth can be stored differently among organizations.

In order to allow for analyses between government organizations, the name of a person will often be the key identifier between records. Between input parties, there must first be an agreement stage where all parties agree on their respective data fields to include and the format they will be provided in. Agreeing on a shared data schema and query occurs before any data submission or secure computation.

6.4 Query Analysis and Computation

After agreeing on a shared schema and preparing their respective datasets, input parties begin by hashing necessary attributes from their records and upload these hashes to the servers in a secret-shared manner. The function (i.e. query) of interest that is to be computed over the data is also submitted. The protocol is based on additive secret sharing, meaning each party partitions their data into three shares and distributes them over servers. These values are generated in such a way that they remain secret-shared among the servers and no server actually learns them [10]. Upon the completion of these uploads, computation is run on the collected data, based on the query provided. Servers jointly compute the function algorithm and give to each result party the aggregated results.

6.5 Sample Query

To give an example of how this architecture can be used, a fictional scenario will be explored in which two government parties collaborate to answer a simple query related to the social determinants of health. This scenario involves a government health data collector that has access to patient health records, and a government tax authority that has access to income records. These parties intend to securely share their datasets to evaluate the query *"What percentage of people who earn under $30,000 a year suffer from heart disease?"*. Heart-related problems are an indicator that there could be limited access to healthy food due to income limitations and this query could be used to determine if governments should enact new policies or expand current ones to support the healthy eating of those in low income brackets.

To begin, both government parties agree on a data scheme and query. Input parties hash their patient records and upload them in a secret shared manner across servers. First, the healthcare data must be transformed where each row corresponds to a single person's list of health complaints. Next, the income data must similarly be transformed where each row corresponds to a single person's income data in the last year. Proof of a privacy-preserving aggregation procedure is outlined in Algorithm 1 of [24]. Finally, these two tables need to be joined by a given person's name and the number of people who both earn under $30,000 per year and have a heart-related complaint need to be summed. The data operations of this query are outlined in Figure 6.

The results of this query will be a percentage value, which will be returned to the necessary parties. This would most likely be both government ministries who contributed their data or potentially another research party that is interested in the results.

6.6 Recommended Systems

Experts expressed concerns over the trust and explainability of MPC, as well as the speed to implement it. Opting for a trusted platform like Sharemind [25] or Senate [31], which have been designed for similar applications, pose the greatest potential of speeding up the adoption process. Sharemind has been utilized for similar queries and use-cases (see Section 3), such as combining and analyzing health, education, and tax data. Both support SQL queries, with Sharemind having an

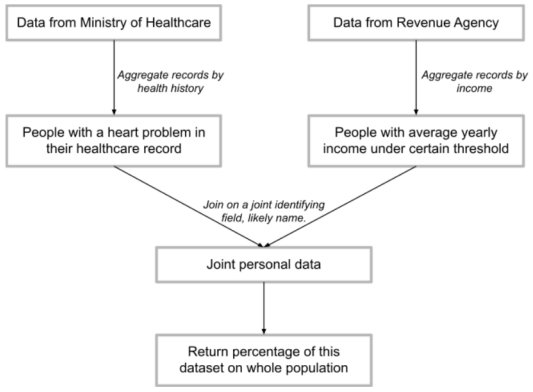


Figure 6: Data operations for proposed architecture on the query *"What percentage of people who earn under $30,000 a year suffer from heart disease?"*

SQL engine and Senate utilizing secure MPC decomposition to efficiently evaluate SQL queries.

6.7 Extension: A Hybrid Approach

While secret sharing is generally more efficient, garbled circuit protocols have seen recent advancements and can be faster for certain queries [32]. A further extension is to create a hybrid MPC architecture that can switch between garbled circuits or a 3-party additive secret sharing approach. This architecture would accept any SQL query and after analyzing the query, determine which approach will be carried out based on estimated execution time. A hybrid approach that utilizes both in parallel has been developed, but not a platform that can dynamically switch on runtime [33]. However, in real-world scenarios, this approach could potentially add unnecessary complexity.

7 Discussion

With the majority of survey participants stating they believe governments should do more to investigate the social determinants of health, the need to improve social health research is evident. All participants agreed both that information on these determinants is important for government systems and that combing data across sectors opens the door for larger investigations to be conducted. While this survey data verified the need for collaboration between government sectors, it also confirmed the feasibility of MPC to support such research endeavours. Additionally, MPC was reported to be a better alternative to current methods of investigating the social determinants of health. These initial findings substantiate the claim that MPC could play an important role in discovering and absolving health disparities.

However, it is important to note that this data reflects only the opinions of ten experts, seven of whom were already familiar with MPC. The lack of MPC implementations in industry is often attributed to the general lack of awareness of this concept, but these survey findings not only state the opposite, but suggest other barriers are responsible. Legislative delays were overwhelmingly reported as a barrier to adopting an MPC-based architecture for social health research across government sectors. Lack of financial support was listed as another key factor, indicating that streamlining regulations to ease the adoption process and justifying why an MPC solution is a necessary investment are the first hurdles to overcome in realising an architecture. At an organization level, participants expressed that there exists a lack of government motivation and perceived necessity to adopt an MPC solution.

While most participants were aware of MPC, distrust in this technology and its explainability were also reported as barriers. This implies that industry *understanding* of MPC is just as relevant as its awareness. To improve trust in MPC implementations, an existing platform like Sharemind or Senate could be utilized. While both systems are off-the-shelf architectures that can ease the adoption process, Senate is far newer (2020) and has not been proven with real government data to the extent that Sharemind has. Senate is also not optimized for all SQL queries, making it less efficient for some computations. However, Senate allows for protection against malicious parties, which could be a requirement in some investigations. As outlined in Section 6.7, leveraging the power of both systems is an avenue for innovation. Utilizing an existing platform can potentially ease the transition to MPC, as there are numerous cases of Sharemind, in particular, being used for similar government applications. An interested government could even collaborate with Cybernetica, the company Sharemind belongs to, to install a solution. Collaborating closely with institutions to adapt a solution for their infrastructure and technical capabilities offers the most potential for success and timely adoption.

While leveraging an MPC architecture shows potential in both discovering new connections between disparate datasets and in providing strong evidence of the scope of known social health disparities, it solves only a portion of the challenges researchers face. A question repeatedly brought up in surveys and interviews is: *Are we storing the right data on these socio-economic determinants?*. In some systems, this may not be the case. Where data is sparse, more health or socio-economic data (new fields, etc.) potentially needs to be collected in order for a researcher's questions to be answered. This approach also only works if the data needed to evaluate a query can be provided and if the correct queries are used. In order to address the structural discrimination that perpetuates health inequalities, the insights gained from MPC need to be used to form new strategies and policies that can truly help those discriminated against. The architecture outlined in Section 6 is dependent on a system that both records necessary information and further leverages the insights provided by MPC research.

Another area of consideration is the need for new data standards. The results of MPC need the ability to be verified for government use, so standards need to be set as to how to elicit and publish verifiable results. MPC makes this extremely difficult, so perhaps sufficient testing and new practices in data pre-processing to prevent corruption and human error could aid in adding robustness to this process. One opinion brought forth by a survey participant is using MPC to enable greater collaboration between private research institutions to increase the accessibility of knowledge. If collaboration ex-

pands, standards and robust practices to ensure the legitimacy of results become even more important. Such strategies and safety nets are necessary to enable greater access to protected data sources.

A key aspect to consider about the proposed architecture is security. Due to legislation, classic means of sharing data are prohibited due to their lack of security guarantees. Current social health studies are often small-scale and even new data legislature can impose bureaucracy rather than technical implementations that achieve higher data security. The architecture proposed in this paper offers higher security guarantees than current methods, but if a government wishes to collaborate with the private sector, where interests could be competing, a semi-honest model may not be sufficient. In this case, other architectures can be explored that offer protection from malicious parties.

An open question of this architecture is scalability; What are the hard computational limits of this proposed solution? Building an MPC architecture for research into the social determinants of health requires designing new technical infrastructure and allocating government resources, making it difficult to begin scaling across many sectors. The limiting network communication complexity of [10] was approximately equivalent to the number of records multiplied by the number of input parties. Multiple implementations listed in Section 3 were able to use MPC for millions of records using Sharemind, supporting the feasibility of leveraging this platform. In cases where such limits are faced, this architecture can still be used for small-scale inter-governmental studies or studies with external researchers.

With MPC protocols, data can ultimately be shared beyond current limits, allowing more research to be securely conducted in this domain. An MPC approach to data sharing can be enacted while policies slowly modernize. This architecture presents a means for research to be conducted in compliance with old policies while we wait for new ones to be enacted. Building secure data bridges between sectors using an MPC architecture presents a new standard of collaboration, one which can unite siloed datasets and leverage collected information to bring about social change.

8 Responsible Research

This research focuses on the sharing of sensitive data, meaning several ethical aspects must be considered to protect individuals should the architecture outlined in Section 6 be implemented. Sufficient testing must be done on the system before utilizing it with real government data. To do so, testing can be done with fabricated data before use.

Furthermore, collaboration with local policymakers needs to be done to ensure that MPC is an acceptable means of data sharing and meets applicable local data security standards. While MPC methods are often compliant with some legislature, such as GDPR, collaboration with local authorities must be done. Data sharing in this secure manner is often a grey area for governments with outdated data policies, so to avoid infringement, this precautionary step must be taken.

Another risk of this architecture is that it could potentially be used to make "super datasets", where all government data could be combined. This presents problems, as storing all data in a centralized location could make it a target for attacks. Thus, this architecture should only be used for collaboration, not for building new permanent super datasets.

Regarding the reproducibility of this research, it should be noted that this paper was informed based on the opinions of experts in the Netherlands, United Kingdom, and Canada. Thus, the need for such a solution is not applicable in countries or government systems where legislature allows for greater data collaboration or where these health determinants are sufficiently investigated. Should the reader be interested in repeating this survey, the questions from the survey are included in Appendix A.1. The architecture proposed has never been created, but following the recommendations of the works in Section 3 offers guidance in doing so.

9 Conclusions and Future Work

Through leveraging a proven MPC platform like Sharemind and optimizing computation through new techniques in data aggregation and block ciphers, the proposed architecture serves as a potential starting point for kickstarting secure privacy-preserving research into the social health domain. Despite barriers such as legislation and funding, it is unanimous among interviewed experts that MPC can open the door for much-needed research that could be leveraged to make health equity a reality. The scalability of this model remains an open question, so a real-world implementation of this idea would need to be developed for further testing. If computation does pose a limitation, MPC could still be feasible for data sharing between sectors or researchers on a smaller scale. Governments recognize this is a problem, but are taking on legislative solutions instead of technical ones. While tackling lack of awareness is a challenge, further advocation of MPC can aid in technical advancements alongside legislative ones.

This architecture stands to show that there exists a need to build bridges between various government sectors. Rather than building a new system from the ground up, we can leverage existing ways of storing data and add a modern, secure data processing framework over it to allow for the sharing of valuable information. There is much work to do on the way to achieving health equity, but being transparent and actively searching for these disparities is a step that needs to be taken.

This paper outlines the industry feasibility of MPC in aiding social health research, but the architecture proposed stands as a solution to a broader problem that is not faced in this domain alone. Extending this approach beyond health data and leveraging it to communicate data between any government organizations is a step beyond this paper. MPC could be used for the broad analysis of any social factor within any industry or field, allowing for social disparities to be analyzed and investigated in other relevant domains. Essentially, this protocol could be used for any dataset that contains personal or confidential information without having to anonymize the dataset ahead of time. New patterns can potentially be found across many sectors of government, which can ultimately lead to sparking progress that brings justice to those systemically discriminated against.

References

[1] Steven A. Schroeder. We can do better—improving the health of the American people. *New England Journal of Medicine*, 357(12):1221–1228, 2007.

[2] Ronald Labonté and Ted Schrecker. Globalization and social determinants of health: Introduction and methodological background. *Globalization and Health*, 3(1):1–10, 2007.

[3] Elissa M. Abrams and Stanley J. Szefler. Covid-19 and the impact of social determinants of health. *The Lancet Respiratory Medicine*, 8(7):659–661, 2020.

[4] Engineering National Academies of Sciences and Medicine. Communities in action: Pathways to health equity. 2017.

[5] Paula Braveman and Laura Gottlieb. The social determinants of health: it's time to consider the causes of the causes. *Public health reports*, 129(1):19–31, 2014.

[6] World Health Organization. *Closing the gap in a generation: health equity through action on the social determinants of health*. 2008.

[7] Ministry of Government and Consumer Services. *Ontario Public Service Data Integration Data Standards*. 2021.

[8] Yehuda Lindell. Secure multiparty computation (MPC). *IACR Cryptol. ePrint Arch.*, 2020:300, 2020.

[9] European Comission. Reform of EU Data Protection Rules 2018. https://ec.europa.eu/info/sites/default/files/data-protection-factsheet-changes_en.pdf, 2018.

[10] Peeter Laud and Alisa Pankova. Privacy-preserving record linkage in large databases using secure multiparty computation. *BMC medical genomics*, 11(4):33–46, 2018.

[11] Andrei Lapets, Frederick Jansen, Kinan Dak Albab, Rawane Issa, Lucy Qin, Mayank Varia, and Azer Bestavros. Accessible privacy-preserving web-based data analysis for assessing and addressing economic inequalities. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–5, 2018.

[12] Hyunghoon Cho, David J. Wu, and Bonnie Berger. Secure genome-wide association analysis using multiparty computation. *Nature biotechnology*, 36(6):547–551, 2018.

[13] Dinusha Vatsalan, Ziad Sehili, Peter Christen, and Erhard Rahm. Privacy-preserving record linkage for big data: Current approaches and research challenges. In *Handbook of Big Data Technologies*, pages 851–895. Springer, 2017.

[14] Vassilios S. Verykios and Peter Christen. Privacy-preserving record linkage. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(5):321–332, 2013.

[15] Thilina Ranbaduge, Peter Christen, and Rainer Schnell. Secure and accurate two-step hash encoding for privacy-preserving record linkage. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 139–151. Springer, 2020.

[16] Rob Hall and Stephen E. Fienberg. Privacy-preserving record linkage. In *International conference on privacy in statistical databases*, pages 269–283. Springer, 2010.

[17] Rainer Schnell, Tobias Bachteler, and Jörg Reiher. Privacy-preserving record linkage using bloom filters. *BMC medical informatics and decision making*, 9(1):1–11, 2009.

[18] Hillol Kargupta, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar. Random-data perturbation techniques and privacy-preserving data mining. *Knowledge and Information Systems*, 7(4):387–414, 2005.

[19] Mehmet Kuzu, Murat Kantarcioglu, Elizabeth Ashley Durham, Csaba Toth, and Bradley Malin. A practical approach to achieve private medical record linkage in light of public resources. *Journal of the American Medical Informatics Association*, 20(2):285–292, 2013.

[20] Joseph I. Choi and Kevin RB Butler. Secure multiparty computation and trusted hardware: Examining adoption challenges and opportunities. *Security and Communication Networks*, 2019, 2019.

[21] Juan Garay, Yuval Ishai, Rafail Ostrovsky, and Vassilis Zikas. The price of low communication in secure multiparty computation. In *Annual International Cryptology Conference*, pages 420–446. Springer, 2017.

[22] Toon Segers. Simplified GDPR compliance using MPC cryptography. https://rosemanlabs.com/article-simplified-gdpr.html, Sep 2020.

[23] Liina Kamm. Privacy-preserving statistical analysis using secure multi-party computation. *Ph. D. dissertation*, 2015.

[24] Dan Bogdanov, Liina Kamm, Baldur Kubo, Reimo Rebane, Ville Sokk, and Riivo Talviste. Students and taxes: a privacy-preserving study using secure computation. *Proceedings on Privacy Enhancing Technologies*, 2016(3):117–135, 2016.

[25] Dan Bogdanov, Sven Laur, and Jan Willemson. Sharemind: A framework for fast privacy-preserving computations. In *European Symposium on Research in Computer Security*, pages 192–206. Springer, 2008.

[26] Claudio Orlandi. Is multiparty computation any good in practice? In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5848–5851. IEEE, 2011.

[27] Dan Bogdanov, Margus Niitsoo, Tomas Toft, and Jan Willemson. High-performance secure multi-party computation for data mining applications. *International Journal of Information Security*, 11(6):403–418, 2012.

[28] Andrew Chi-Chih Yao. How to generate and exchange secrets. In *27th Annual Symposium on Foundations of*

Computer Science (sfcs 1986), pages 162–167. IEEE, 1986.

[29] Debra Grant. *A Guide to the Personal Health Information Protection Act–Rev*. Information & Privacy Commissioner of Ontario, 2004.

[30] Chun Guo, Jonathan Katz, Xiao Wang, and Yu Yu. Efficient and secure multiparty computation from fixed-key block ciphers. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 825–841. IEEE, 2020.

[31] Rishabh Poddar, Sukrit Kalra, Avishay Yanai, Ryan Deng, Raluca Ada Popa, and Joseph M. Hellerstein. Senate: A maliciously-secure MPC platform for collaborative analytics. *CoRR*, 2020.

[32] Aner Ben-Efraim, Yehuda Lindell, and Eran Omri. Efficient scalable constant-round MPC via garbled circuits. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 471–498. Springer, 2017.

[33] Pille Pullonen and Sander Siim. Combining secret sharing and garbled circuits for efficient private IEEE 754 floating-point computations. In *International Conference on Financial Cryptography and Data Security*, pages 172–183. Springer, 2015.

A Appendix

A.1 Survey Questions

The format of this survey was a Google Form. For all questions, participants had the option to answer the question or not. All participants answered the non-open questions and a further few also answered the open questions.

Introductory Questions

Thank you for participating in the following survey on multiparty computation (MPC) and the social determinants of health. This survey takes approximately 10 minutes to complete and both MPC and the social determinants of health will be explained along the way! The results will be used anonymously to aid my Bachelor's student thesis.

Social factors often have more influence on physical and mental health than clinical factors. These are known as the social determinants of health and include factors like a person's race, level of education, income, and the location they live in. An example is that someone with a lower income level could be limited to eating unhealthy food options, ultimately resulting in an increased risk of obesity or heart disease.

There will be an opportunity to share your comments in every section.

For any questions or concerns, please feel free to contact me (Brontë Kolar) at b.t.a.kolar@student.tudelft.nl.

1. Please check the industry you are involved in:
    - Healthcare
    - Legal (law or policy)
    - Data privacy and security
    - Other *(participant can add a custom field)*
2. What is your job title?
3. What is your name? This is optional and would be used to list you as a contributor in this thesis paper. It will not be tied to your responses.
4. Which country do you live in?
    - Canada
    - The Netherlands
    - United States
    - Other *(participant can add a custom field)*
5. Are you familiar with the social determinants of health?
    - Yes
    - No

Social Determinants of Health

While these social determinants of health have been investigated, we are currently limited with the scale they can be investigated on.

Determining how social determinants affect healthcare on a large scale involves gathering data from different siloed government institutions and combining it with healthcare data. These institutions are often not legally allowed to share data with each other and there are strict regulations on the sharing of healthcare data, which is considered very sensitive. In current investigations, personal privacy and security are top of mind for industry stakeholders.

While not easy to do with the current legislature, taking steps to combine these siloed datasets could aid in the investigation of these determinants (unveiling new links and trends, etc.).

A useful example of this would be combining data from a national revenue agency with patient healthcare data to determine if citizens with lower income levels suffer greater from certain illnesses. This information can then be leveraged in policy and healthcare. For instance, governments could enact policies that target early screening for people in certain income brackets. Currently, this type of data sharing is very limited by legislature.

While it is possible to collect this data through a consent-driven survey, the scale is smaller and ensuring the information is representative of whole patient population presents another barrier.

1. Do you believe governments should do more to investigate social determinants of health?
    - Yes
    - No
2. Which barriers stand in the way of investigating social determinants of health?
    - Legislature
    - Technology
    - Willingness to investigate these determinants
    - Other *(participant can add a custom field)*
3. If government organizations shared more data between each other, do you believe larger investigations into social determinants of healthcare could be conducted?
    - Yes
    - No
4. Do you believe more information on the social determinants of health is important for healthcare and government systems?
    - Yes
    - No
5. Feel free to add any comments on your responses here (*open response*).

Multiparty Computation (MPC)

MPC is a set of protocols that enables multiple parties to submit data securly and perform computations on said data, such that each party learns nothing beyond its own input. Data is fully secured during computation and each party is fully unaware of the input data of other parties.

Here's an example to help visualize this: If we were both millionaires and wanted to know who was richer without revealing our salaries to each other, we could use an MPC protocol. We could securely submit our salaries and the protocol would output only who is richer, without ever revealing what the submitted salaries are. We would have no idea what we submitted.

MPC makes it possible to analyze massive datasets without ever revealing sensitive information about the data, allowing

for privacy-preserving statistics. This can be used to investigate the social determinants of health because healthcare and social government institutions can submit their data securely using an MPC protocol. Ultimately, a secure statistical analysis on the data can be run and new insights on these social determinants can be unveiled.

An example could be the following: A database of healthcare data can be combined with a database of revenue data using a secure MPC protocol. The protocol could evaluate the percentage of people who make less than $30,000 a year and suffer from heart disease. The MPC protocol would then return just this percentage. Thus, during and after the process, no patient or revenue data would ever be exposed!

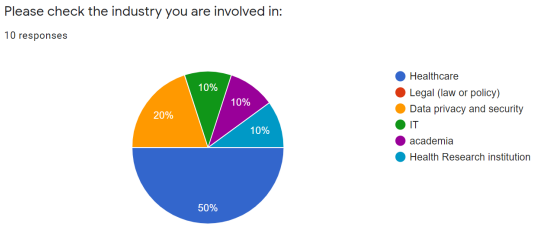
1. Are you familiar with MPC?
   - Yes
   - No
2. Do you think using MPC could be used to combine data across institutions to investigate the social determinants of health?
   - Yes
   - No
3. Do you think using MPC to combine data across institutions to investigate the social determinants of health is a feasible idea (*1 = not feasible, 5 = very feasible*)?
4. Do you think MPC is a better alternative solution to current methods of investigating these determinants (*1 = worse than current methods, 5 = better than current methods*)?
5. What barriers do you foresee stand in the way of adopting this technology (*multi-select*)?
   - I see no barriers
   - Distrust in technology
   - Concern that it is too new
   - Lack of financial support
   - Legislative / regulatory delays
   - Other *(participant can add a custom field)*
6. Why do you think organizations are not already using this technology (*multi-select*)?
   - Lack of awareness
   - Lack of government motivation
   - Lack of funding
   - This technology is not perceived to be necessary
   - Other *(participant can add a custom field)*
7. Feel free to add any comments on your responses here (*open response*).

Final Thoughts

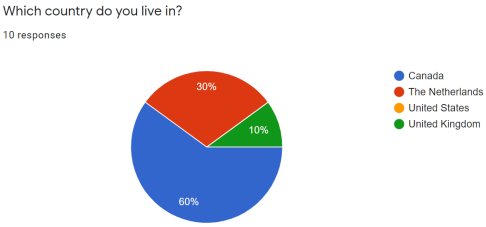
1. What extra considerations or requirements do you think need to be taken into account for an implementation of MPC to investigate social determinants of health (*open response*)?
2. Do you see any other potential use-cases for MPC (*open response*)?
3. Please note any other thoughts or ideas you would like to contribute to this survey (*open response*).
4. If you would like to receive a digital copy of my final thesis paper on this topic, please leave your email below (*open response*).

A.2 Survey Responses

Introductory Questions



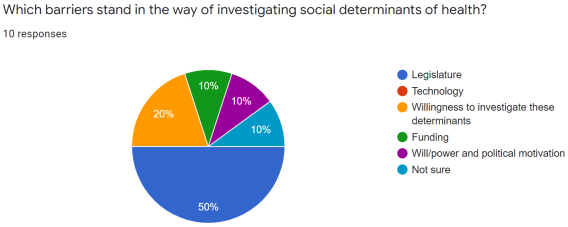
1.
2.
   - Senior Privacy Specialist
   - Director, Project Management Office
   - Data Scientist
   - Researcher
   - SR. Manager CyberSecurity
   - Manager, Strategy
   - Computer Science and Engineering student
   - Director, Architecture
   - associate professor
   - Director, Data Quality and Information Management
3. Responses omitted for anonymity.



4.
5. 8 - Yes, 2 - No

Social Determinants of Health

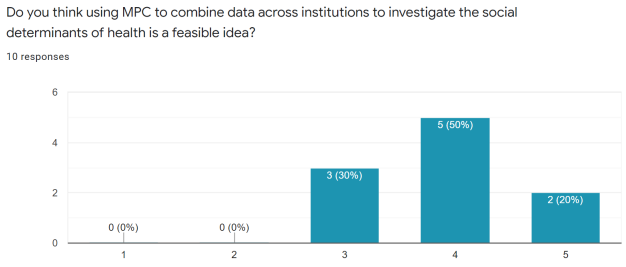
1. 9 - Yes, 1 - No

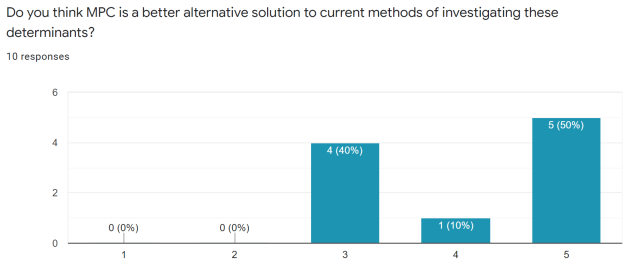


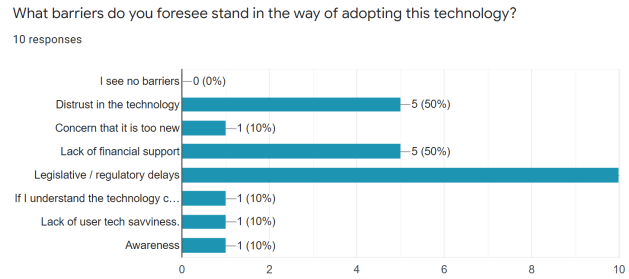
2.
3. 10 - Yes, 0 - No
4. 10 - Yes, 0 - No
5.
   - Anonymisation of healthcare data is required to share it.

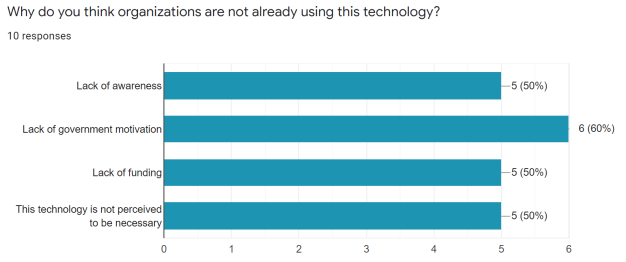
- More socio-economic data needs to be collected then integrated back with healthcare, housing, education data etc. . .
- In most of the yes/no questings I felt like answering "I don't know" would have been a better response than choosing between one of them, since I don't know a lot about social deterninants (and therefore also do not really have an opinion on some of the question).

Multiparty Computation (MPC)

1. 7 - Yes, 3 - No
2. 10 - Yes, 0 - No
3. 

4. 

5. 

6. 

7. - This would make my work of combining sensitive data from different sources much easier. However, working with some organizations that would have to implement/support this into in their data environment makes me skeptical that it is feasible in a timely manner.If this would be offered as a third-party service with all the regulatory permissions checked already I believe it would speed up the process of adoption significantly.
   - A standard also needs to be set as to how to publish verifiable results

Final Thoughts

1. - Modernization of health privacy legislation and a strong commitment to funding MPC implementation.
   - MPC should be combined with Open Science to enable greater collaboration between institutions. Standards need to be created to enable greater access to anonymized data sources.
   - Data sovereignty, lifecycle controls
   - Privacy and potential for bias in the way code/algorithms are developed / utilized
   - Are we storing the right data to make these determinations.
2. - No at this time.
   - I think that MPC could be used for broad analysis of any social factor within any industry or field, not just healthcare. Essentially you could use this protocol for any dataset that contains personal or confidential information without having to anonymize/sanitize the dataset ahead of time.
   - Crypto trust protocols and oracles.
   - Socioeconomic barriers to healthcare. Privacy by design
   - Housing is a major concern across metropolitans, data can help us solve and prevent it through better planning and socio-economic safety nets.
   - Dating apps (doesn't Tinder do something like this, for showing matches?)
3. - Challenges of eliminating bias when deriving results from the dataset.
4. Responses omitted for anonymity.