Workshop on privacy-preserving statistical computation with Statistics Norway

Makhlysheva A., Yigzaw K.Y., Nordsletta A.T., Bellika J.G.
Workshop on privacy-preserving statistical computation with Statistics Norway

Report number
07-2018

Project manager
Alexandra Makhlysheva

Authors
Alexandra Makhlysheva
Kassaye Yitbarek Yigzaw
Anne Torill Nordsetta
Johan Gustav Bellika

ISBN
978-82-8242-089-1

Date
31.05.2018

Number of pages
24

Keywords
Personal data, health data, privacy, security, confidentiality, patient-generated data, statistics

Summary
The problem of privacy preserving for distributed computations is relevant in healthcare and other domains where statistical analysis of data combined from several sources is performed.

With secure multi-party computation, all the computations are performed without revealing any microdata to the external entities: each party learns only the corresponding function output value and no inputs of other parties. This is especially useful when the source databases are confidential and should not be openly linked.

Both individuals’ and institutions’ privacy are protected without affecting the quality of research results.

Publisher
Norwegian Centre for E-health Research
PO box 35, NO-9038 Tromsø, Norway

E-mail: mail@ehealthresearch.no
Website: www.ehealthresearch.no
Table of contents

Summary ..................................................................................................................................... 4

1 Background ................................................................................................................................ 5
  1.1 BigInsight .................................................................................................................................. 5

2 Arrangement .................................................................................................................................. 6
  2.1 Day 1. Open Seminar on Privacy-Preserving Distributed Statistical Computation ........... 6
    2.1.1 Stein Olav Skrøvseth: Norwegian Centre for E-health Research ................................. 6
    2.1.2 Øyvind Langsrud: Modernization of Statistics Norway - 2022 ....................................... 10
    2.1.3 Johan Heldal: RAIRD in five minutes! .............................................................................. 12
    2.1.4 Giske Ursin: Health data and the re-identification threat – a real world example ...... 12
    2.1.5 Kassaye Yitbarek Yigzaw: Distributed data analysis in the face of privacy concerns... 15
    2.1.6 Johan Gustav Bellika: The Norwegian Primary Care Research Network IT infrastructure: The Snow system .............................................................................................. 15
    2.1.7 Kassaye Yitbarek Yigzaw: Privacy-preserving collection and analyses of citizens-generated data .......................................................................................................................... 18
  2.2 Day 2. Discussion between NSE and SSB on further collaboration ..................................... 19

3 Discussion ..................................................................................................................................... 21

4 References ................................................................................................................................... 23

List of figures

Figure 1. One citizen - one health record .......................................................................................... 7
Figure 2. Learning Healthcare System ............................................................................................. 8
Figure 3. Big data in healthcare. IBM view ....................................................................................... 9
Figure 4. Gartner Hype Cycle for Emerging Technologies, 2017 ................................................. 10
Figure 5. Future innovation environment at SSB .......................................................................... 11
Figure 6. ARX re-identification risk analysis ................................................................................ 14
Figure 7. The Snow system architecture. Coord = Snow Coordination server, S = Snow Server in local health institution ................................................................. 16
Figure 8. Data flow in PCRN ........................................................................................................ 17
Figure 9. Distributed statistical computations with Emnet/Snow ................................................ 18
Figure 10. Secure data collection .................................................................................................... 19
Summary

Based on the presentations by the speakers from the Cancer Registry of Norway, Statistics of Norway and Norwegian Center for E-health Research, as well as the number of questions raised up from the audience, we can conclude the problem of privacy preserving to be relevant both in healthcare and other domains where statistical analysis of data combined from several sources is performed.

Different solutions can be applied for this issue. Pseudonymization and de-identification techniques still leave the space for sneaking into the individuals’ data due to more information available in the joined databases; linking additional publicly available information about the individuals (for example, from social media) enhances the privacy risk.

Secure multi-party computation techniques based on blind data miners can become a solution for preserving privacy in statistical studies with data from several sources. With SMC, all the required computations are performed without revealing any microdata to the computing entities: each party learns only the corresponding function output value and no inputs of other parties. The blind data miners run secure protocols that compute statistical functions on the data producing the aggregated results. This is especially useful when the source databases are confidential and should not be openly linked. Herewith, both individuals’ and health institutions’ privacy can be protected saving the quality of research results. Additionally, costs of conducting statistical studies on distributed data can be reduced.
1 Background

Providing good healthcare to individuals requires storing and using health-related information. Privacy in healthcare sector deals with the patient’s autonomy, justifiable processing of health-related data, together with correct health data provided to the right healthcare provider. According to the article 9 of Declaration of Helsinki, “it is the duty of physicians who are involved in medical research to protect the life, health, dignity, integrity, right to self-determination, privacy, and confidentiality of personal information of research subjects” [1]. Ensuring health information to be not sent to unauthorized persons is fundamental to form trust between patients and healthcare professionals. Trust is a prerequisite for providing quality healthcare. In addition, research and statistical institutions processing health-related data must do it in a proper, privacy-preserving manner so that this information is not compromised. These principles are applicable to both Norwegian Center for E-health Research (NSE) and Statistics Norway (SSB).

Statistics Norway is the national statistical institute of Norway responsible for collecting, producing and publishing official statistics related to the economy, population and society at national, regional and local level. SSB also performs extensive research and analytical activities. Most of Statistics Norway’s analytics are based on data from administrative registers and surveys. Additionally, SSB has a growing interest in reuse of information coming from different business actors. Privacy and data security issues are central in SSB. SSB is subject to the Personal Information Act regarding the proper processing of personal data [2]. SSB recently joined the BigInsight consortium (read more about BigInsight further in the text).

NSE has been doing research on techniques for privacy-preserving analysis of distributed health data. Through exchanging NSE’s experience and how SSB treats aggregated population data, both institutions can enrich their own toolboxes for maintaining privacy in the processing of personal data as well as establish contact for further collaboration projects.

1.1 BigInsight

BigInsight is a research-based innovation center funded by the Norwegian Research Council and 15 public, private and research consortium partners [3]. The consortium consists of the following partners: Norwegian Computing Center (Norsk Regnesentral), University of Oslo, University of Bergen, ABB, DNB, DNV GL, Gjensidige, Hydro, Telenor, NAV, Skatteetaten, Oslo University Hospital, Norwegian Institute for Public Health, Cancer Registry of Norway, and Statistics Norway.

The objective of the BigInsight is to work with statistics and machine learning for performing predictions and decisions [4]. The consortium aims for developing new statistical tools for performing analytics of high dimensional and complex data that will be used in personal marketing, personalized health and patient safety, personalized fraud detection, sensor systems, and forecasting power systems [4].

The researchers are developing new mathematical, statistical and computational methodology for improving breast cancer predictions. In addition, they are developing methods for foreseeing synergy between drugs or the effect of drugs combination using data from cancer cell lines and models for predicting cancer drug sensitivity with large-scale in vitro drug screening [5].

BigInsight uses the following methodologies and techniques in its projects: clustering, focused inference, functional data analysis, graphical models, hierarchical Bayesian models, data integration, model comparison and model improvement, multiple testing, multivariate dependence and copula models, extreme value theory, non-parametric Bayes, non-stationary and non-linear stochastic processes and time series, sequential inference, stochastic geometry and space time models, subsampling and data thinning [4].
2 Arrangement

SSB and NSE arranged a two days seminar on March 05 and 06 to discuss privacy and confidentiality challenges and possible solutions, and plan for future collaborations.

2.1 Day 1. Open Seminar on Privacy-Preserving Distributed Statistical Computation

On the first day, BigInsight and Statistics Norway organized a half-day seminar on privacy-preserving distributed statistical computation. The seminar was open for public.

NSE was represented by Kassaye Yitbarek Yigzaw, Johan Gustav Bellika, Anne Torill Nordsletta, Stein Olav Skrøvseth, and Alexandra Makhlysheva.

Stein Olav talked about NSE, its role, and a role of data in e-health research. Kassaye Yitbarek Yigzaw had two presentations entitled “Distributed data analysis in the face of privacy concerns” and “Privacy-preserving collection and analyses of citizen-generated data”. Johan Gustav Bellika presented “The Norwegian Primary Care Research Network IT infrastructure: The Snow system”.

Other speakers on the seminar were:
- **Jørn Leonhardsen**, IT Director, responsible for IT, data capture and statistical method work in addition to the professional responsibility for developing common functions in preparation and analysis in Statistics Norway’s statistics production. He is also a steering board member at BigInsight
- **Øyvind Langsrud** and **Johan Heldal**, researchers from Section for methods of Department of Digitalization and Joint Functions in Statistics of Norway
- **Giske Ursin**, Director for the Cancer registry of Norway

Jørn Leonhardsen opened the seminar by denoting the benefits and problems of data sharing. Data sharing involves risk with jeopardizing privacy preserving. Valuable knowledge can be obtained by combining data from two or more sources, but exchanging and linking data are often unacceptable due to confidentiality and privacy concerns. Consequently, important for society discoveries could be hampered. Therefore, data and results need to be processed in a privacy-preserving way. New approaches and computing methods for analyzing data distributed across multiple data sources while protecting privacy are being developed.

2.1.1 Stein Olav Skrøvseth: Norwegian Centre for E-health Research

Norwegian Center for E-health Research (NSE) was established 1 January 2016. It is organized as a research center at the University Hospital of North Norway. NSE has around 60 employees, including 7 PhD students, 20 researchers, 15 advisors, 8 professors, and of these around 30 have PhDs. The center has competence in medicine, technology, social sciences/sociology, and economics.

NSE replaced Norwegian Centre for Integrated Care and Telemedicine (NST) that, according to the Ministry of Health, had fulfilled its mission as a competence center for telemedicine. The Ministry of Health wanted to use the existing competence of NST in a new context. The Directorate for e-health was established at the same time to implement the national policy on e-health, establish the requisite standards, and administrate the use of e-health methodology nationwide. In the National Budget 2017, it stated, “the national center (NSE) will support national needs for research and analysis in e-health, telemedicine and mobile health, including evaluation research and health technology assessment of e-health.” The political aim of the center is to support the goals of “One citizen – one health record” which is described in the white paper 9 (2012–2013) (Figure 1).
The national projects are cooperation projects with national institutions. In 2018, the projects are thematically aligned with:

- “Helseplattformen” in Helse Midt RHF
- Modernization of the electronic health record (EHR) in the other health regions
- One citizen – one health record
- Welfare technology

NSE has four strategic areas and departments:

- Future EHR
- Integrated services
- Personal systems and connected care
- Availability and analysis of data

The center has a significant research activity financed through competitive funds, for example, the Research Council of Norway and the EU. This activity will support the national mission and ensure the center has strong and competitive professional competence on the topics relevant to national developments in the e-health field.

The National Center for E-Health Research continues its cooperation with the World Health Organization (WHO), which has been in progress with the National Center for Telemedicine since 2002. The new center is a member of the EHTEL (European Health Telematics Association) and HIMSS Governing Council Europe. The research community has also extensive experience from participation in EU projects and international research networks. In cooperation with the Directorate for e-health, the Center will help to organize and coordinate Norwegian participation in various international arenas.

For improving the quality of healthcare services, the new approaches for extracting knowledge from data are needed. Learning healthcare system (LHS) (see Figure 2) is the process of generating knowledge from data coming from health services and operations and applying the knowledge into clinical practice. The realization of this objective depends on health information technology, such as databases, EHR, and research infrastructure.
The amount of data created and stored on a global level is almost inconceivable and keeps growing. In his book Megatrends (1982), John Naisbitt wrote, “We are drowning in information but starved for knowledge.” It is relevant now, more than ever. Companies have a lot of accumulated data. Having too much data makes it difficult to understand which data to focus on. Additionally, despite having huge amounts of data to access, many companies cannot draw the important insights they need to turn data into actionable information. Insights from big data can lead to better decisions and strategic moves. Herewith, big data is coming from several sources, has quality concerns, is inherently biased, unstructured, and somewhat “complete”. Big data in healthcare has the same challenges (Figure 3).
The use of IBM supercomputer in cancer treatment, which has still not reached the IBM’s promises and customers’ expectations, is an example of unsolved big-data related issues in healthcare.

Data analysis in healthcare can be successfully used for:

- Free text analysis to give prediction of the patient’s treatment
- Decision support for clinicians and patients
- Prevent unwanted events
The international consulting company Gartner provides a graphic representation of the maturity and adoption of technologies and applications, and how they are potentially relevant to solving real business problems and exploiting new opportunities. Gartner considers AI as a mega-trend, which will be the most pioneering technology for the next five to ten years [7] (Figure 4).

![Gartner Hype Cycle for Emerging Technologies, 2017](Figure 4. Gartner Hype Cycle for Emerging Technologies, 2017. Source [7])

AI is about to be used in technology services and applications, also in health sector. NSE recently published a health analytics report with a focus on machine learning, natural language processing, data mining and process mining methods, including usefulness of these technologies, use cases, tools and relatedness to Norwegian healthcare [8].

2.1.2 Øyvind Langsrud: Modernization of Statistics Norway - 2022

Statistics Norway plans to modernize its statistics production through a comprehensive digitization program, exploitation of new skills, and use of new technologies and methods to support their societal mission [9].

In order to analyze more data, make versatile comparisons, streamline data retrieval, and automate manual routines, Statistics Norway plans to produce statistics in a different way in the near future. The modernization program involves investment in new technology and increased technological competence [9].

Registers and sample services are traditional data sources for Statistics Norway. They are involved in the long-term modernization program, which also includes modernization of KOSTRA (municipality state reporting).

In the near future, SSB is going to involve conventional sources, big data sources coming from several providers to its analyses. These data sources include transaction data, such as data coming from pay-
ment transactions, cash registers/scanners, memberships, satellites, mobile positioning, and transportation (Figure 5). SSB is involved in international collaboration with Eurostat supported projects (ESSNet) and has partnership with Center for Big Data Statistics, Statistics Netherland.

SSB’s confidentiality protection measures include table protection, microdata protection, and RAIRD.

Table protection in SSB is done by cell suppression and rounding. In frequency tables, small values cannot be shown (primary suppression), and additional cells must be hidden (secondary suppression). In quantitative tables, cell suppression is done by other rules followed by secondary suppression.

SSB also uses rounding for protecting frequency tables following such rules as:

- Changing small values
- Changing only necessary small values in the inner table (all variables are cross-classified) so that values in the tables to be published (selected cross-classifications) are safe (this method was created by Johan Heldal)
- Making sure of important sums/totals to be published do not change significantly

SSB implemented these table protection methods as an R package following a specified input/output standard. This package is embedded within the IT system at SSB. Several of the functions was implemented within the KOSTRA project (this includes among others table suppression and rounding).

SSB plans to include transaction data in future production. In this context, they have several issues to deal with. These issues are:

- New statistics by combing transaction data with other sources (population register)
- Is storing raw transaction data inside the SSB needed/accepted?
- Is it possible to compute statistics without exchanging and linking data?
2.1.3 Johan Heldal: RAIRD in five minutes!
RAIRD is Remote Access Infrastructure for Register Data. It is a query system for statistics and analysis with register data. It is a cooperation between SSB and Norwegian Centre for Research Data (NSD) in Bergen, and funded by NFR (Norwegian Research Council) since 2012. The system can be accessed only via Internet. It is targeted for researchers in approved institutions. The system is now on the testing phase. There are plans to open the system for general use in the near future.

There are several registers involved in the RAIRD project:

- The general population register
- NAV’s register
- Incomes and taxes (Skattedirektoratet, the Norwegian Tax Administration)
- Employer-employment registers (A-ordningen)
- Education (NUDB, the National Education Database)

Access to the system is possible only through a web page via a signed agreement and logging with ID-porten.

A query system of RAIRD works as follows:

- No microdata leaves SSB
- Only statistical output to users – no microdata
- Output is subject to statistical disclosure control

The statistical disclosure control is implemented by the following measures:

- Limitations on populations sizes (output should have at least 1000 people)
- Group disclosure
- Noise on all counts (constant random noise based on record key and cell keys technology)
- Excluding dominants
- Differencing (subtracting tables for almost equal populations)
- Graphic displays are “coarsened” using Hexbin plots.

A Hexbin plot is useful to represent the relationship of two numerical variables when there are a lot of data point. Instead of overlapping, the plotting window is split in several hexbins, and the number of points per hexbin is counted. The color denotes the number of points.

2.1.4 Giske Ursin: Health data and the re-identification threat – a real world example
The Cancer Registry of Norway has access to 17 central health registries and 54 clinical registries in Norway. They also have data from different biobanks and population surveys. All these data can be linked. The Cancer Registry collects data on different types of cancer to assess distribution of disease and obtains information on how to prevent disease and death caused by the disease.

Among others, the Cancer Registry has data on screening for cervical cancer and mammographic screening of all women of age 25-69. These data include the following information about each of approximately one million women who had this test:

- Month and year of birth
- Dates of all cervical exams
- Results of each test
- Whether or not got cancer
- Cancer diagnosis date

Storing these data raises several questions:

- Where is it safe to store the data? It should only be accessed there
- If we link more data, are the data still safe?
• The new health platform coming. Will the data be safe there?
• What about the re-identification threat? The systems available for researchers nowadays are based on trust.

The Cancer Registry of Norway got a request from researchers to get data from the cervical cancer program with all the results from the screening data (5.6 million records, 915 thousand women). The company wanted to have the datasets on every woman in the registry. The Cancer Registry is required to hand out data when people have the reason to perform research. However, they have to check if linking the data with other datasets reveals sensitive information about individuals. To reduce re-identification risk, the Cancer Registry performed K-anonymization of data (categorized variables), fuzzification, and created synthetic datasets. The challenge with synthetic datasets was what to do if some aspect of calendar year needs to be changed.

In the first round of fuzzification, the Cancer Registry altered the data. They did K-anonymization, excluded some observations, and altered all dates: removed days in dates, changed months in dates with a random number, and removed months of birth. In the second round of fuzzification, they removed extreme dates and combinations, changed the id (assigned a new random id), changed all the dates with the same random number, used fuzzy factor on months, set all days in dates to 15.

For assessing the risk of re-identification of individuals in the dataset, the ARX tool was used. It quantifies risk of re-identification based on uniqueness, has a prosecutor scenario by assuming a person in dataset, and classifies variables as identifiable, quasi-identifiable, or sensitive. The researchers checked three datasets: 1) realistic dataset, 2) K-anonymization of the first dataset, and 3) dataset with fuzzified months in the second dataset.

In the original database, the average prosecutor risk was 97.06%, while the average prosecutor risk for the second dataset was 9.7%, and for the third dataset this parameter was equal to 9.8% (Figure 6).
Categorizing date variables (applying K-anonymization) substantially reduced the possibility of re-identification of individuals. Adding a random factor, such as a fuzzy factor, makes it even more difficult to re-identify specific individuals. It was concluded that simple techniques could substantially reduce the risk of re-identification.

Ursin and colleagues published an article where they explained their method presented above in the Cancer Epidemiology, Biomarkers & Prevention journal [10].

The current regulation of health data processing is done by Health Data Registry Law paragraph 6 (Helseregisterloven). However, the registries define how much data they provide. The law is the reason for long time to hand out the data. The new regulation, GDPR (Data Protection Impact Assessment Article 35), has already entered into force. The sufficient information of what has to be done in accordance with the new regulation is awaited from the Norwegian Data Protection Authority (Datatilsynet).

The Norwegian Directorate of e-health is developing a national health analytics platform, which will “… simplify access to health data and facilitate advanced analytics across health registries, source data, health records and other sources of health information.” Giske argues the platform cannot be built on trust alone. There should be some data protection mechanisms. For the researchers, there should be a balance between safe analysis of large linked data (excluding re-identification threat) and rapid and seamless analysis combined with ability to check individual records. Therefore, there is a need for national platforms able to provide it all.
2.1.5 Kassaye Yitbarek Yigzaw: Distributed data analysis in the face of privacy concerns

The increased adoption of electronic health record systems, as well as a wide variety of other electronic data sources (for example, insurance claims and registry data), led to the collection of large amounts of detailed health information about individuals. Reuse of these data has huge potential for a variety of purposes, such as research and public health. Data reuse increases the rate of new scientific discoveries and answers research questions that may not be possible otherwise.

Generalizability and reproducibility of analysis results often require data from multiple data sources. The data from a single institution may not have a sample size large enough to provide sufficient statistical power or heterogeneity that represents the population of interest.

Secondary use of data raises privacy concerns on individuals’ level. Inappropriate disclosure of sensitive information may lead to mental and physical harm to patients. Even when individuals’ privacy concerns are addressed, clinicians and healthcare providers are also concerned that data sharing may damage doctor-patient relationship, data could be used to evaluate their performance, or, in some contexts, reveal confidential business information. Therefore, privacy concerns limit willingness for data sharing.

Traditionally, distributed data is centrally collected at a trusted third party who analyses the data. The third party can be an institution like SSB or a research center. The data collected at the third party can be either patient-identifying or de-identified. Reuse of patient-identifying data often requires consent, although it can be exempted in rare cases. A systematic difference between individuals who consent and do not consent, can lead to bias. In addition, consent collection is expensive and takes long time.

Sharing of de-identified data often does not require consent. The main challenge of de-identification is making a balance between re-identification risk and data utility. The problem becomes even more demanding in the context of distributed data. A simple approach for de-identifying distributed data is local data de-identifying by each institution before sharing. However, the union of the de-identified data does not give the same result as de-identification of centrally collected data.

There is an area of research called secure multi-party computation (SMC). It deals with the problem of computation on distributed data without revealing anything apart from the result. The main challenges of SMC are efficiency - an ability to compute with good performance, and scalability - an ability of efficient computing when the number of data custodians and records increases. A tool Emnet [11] has been developed for privacy-preserving statistical computation of data distributed across data custodians. It is efficient and scalable, and enables data custodians to maintain access control on their data.

2.1.6 Johan Gustav Bellika: The Norwegian Primary Care Research Network IT infrastructure: The Snow system

Gustav started his presentation with citations from the Declaration of Helsinki. In particular, he referred to the Articles 6 and 9. “The primary purpose of medical research involving human subjects is to understand the causes, development and effects of diseases and improve preventive, diagnostic and therapeutic interventions (methods, procedures and treatments). Even the best-proven interventions must be evaluated continually through research for their safety, effectiveness, efficiency, accessibility and quality” (Article 6) [1]. “It is the duty of physicians who are involved in medical research to protect the life, health, dignity, integrity, right to self-determination, privacy, and confidentiality of personal information of research subjects. The responsibility for the protection of research subjects must always rest with the physician or other healthcare professionals and never with the research subjects, even though they have given consent” (Article 9) [1]. This implies medical research to be privacy preserving.

Objectives of creating the Norwegian Primary Care Research Network IT infrastructure are
- Make participation in research projects easier and more efficient for the GPs
- Reuse health data in a safe and privacy preserving manner
- Complete research projects according to scheduled time and resources consumption
- Recruit 90-110 GP practices
- Cover 7.5% of the Norwegian population

The Snow system is a distributed system that enables collection and reuse of anonymous medical data, builds and maintains a national online epidemiology model, uses the epidemiology model to provide automated IT-based health services, enables privacy-preserving distributed computations on EHR data, which is directed at research, quality improvements, audit, and disease surveillance.

The Snow system is a “collaborative edge computing” infrastructure performing coordinated computations on distributed resources (Figure 7). Edge computing refers to the enabling technologies allowing computation to be performed at the edge of the network. It is beneficial when data is too sensitive (health data), too big (genetic data), or too competitive (data will expose profile of an owner).

The computing entity in the network is an individual computing process (one instantiation at each participating Snow server) which has a unique communication address. Agents communicate between each other using XMPP messages. Computations are coordinated: there is one “main” agent coordinating “missions” of multiple agents performing computations in parallel. Each node in the network is a Snow appliance box, a small computer with pre-installed Snow server software that can be administrated remotely by the Snow team at UiT/NSE. Both patient and GP’s data in the box is pseudonymized. The schema of data flow in PCRN is given in Figure 8.

![Figure 7. The Snow system architecture. Coord = Snow Coordination server, S = Snow Server in local health institution](image-url)
Further Gustav was talking about possibilities of using secure multi-party computations (SMC) to support research in primary care.

Creating a virtual dataset with Emnet/Snow can help for making participation in research projects easier and more efficient for the GPs, supporting researchers in terms of including of sufficient number of patients in clinical research, and supporting article 9 in Declaration of Helsinki, i.e. being used for privacy preserving. Based on the assumption, the EHRs are using OpenEHR as a common data model. Therefore, the research criteria specified by the researcher are written in archetype query language (AQL) - a query language used to query OpenEHR-based database. The AQL is executed at each hospital to select datasets that fulfils the research criteria. These datasets are stored in a separate database locally. The datasets at each hospitals collectively form the research data referred as a virtual dataset.

Once the virtual dataset is created, the next question will be how to perform privacy-preserving analysis on the dataset. Therefore, the researcher sends another query that contains statistical function and variables to the coordinator server, and the coordinator server and the hospitals perform computations securely using a technique called secure multi-party computation. SMC enables to jointly compute on private data but reveals only the final computation result. An example of SMC is illustrated on Figure 9.
Automated processing at PCRN is aimed at supporting Article 6 in the Declaration of Helsinki, i.e. continuous evaluation of interventions.

The PCRN can beneficially help GPs become more efficient in research has the following benefits through centralized resources as PCRN staff/researchers, can produce aggregated (non-sensitive) statistics automatically directly from the sources, and generate knowledge about the patient populations directly from the distributed sources, spanning administrative borders as municipalities, regions, countries and continents.

However, two other comparable approaches exist, and no standard is established. There is also an issue of how to validate correctness of computed statistics.

2.1.7 Kassaye Yitbarek Yigzaw: Privacy-preserving collection and analyses of citizens-generated data

Due to the recent advances in mobile and sensor technologies, and “quantified self” movement\(^1\), individuals are capturing large volume of health-related data outside the traditional healthcare settings. Moreover, large amounts of data are also being collected from individuals through questionnaires. Citizen-generated data have huge potential for a variety of purposes, such as research and public health. The data increase the rate of new scientific discoveries and enable to answer research questions not feasible otherwise.

The common approach for analyzing citizen-generated data involves central collection of patient-identifying or anonymous data at a third party (i.e. researcher and institution). For anonymous data collection, only direct identifiers are removed. However, re-identification is possible through indirect identifiers.

Individuals are responsible for capturing and recording of the data they generate. As a result, they have direct control over their data. Therefore, if individuals fear for their privacy, they may not be willing to share their data or provide inaccurate information. Systematic difference between individuals who provide their data and those who do not provide can lead to statistical bias. Therefore, data analysis techniques able to address the privacy concerns are needed.

A tool designed and implemented for privacy-preserving collection and analysis of citizen-generated data [12] was discussed further (Figure 10). In this tool, individual’s secret values are split into parts,

---

\(^1\) [http://quantifiedself.com/](http://quantifiedself.com/)
called shares, and each share is sent to a third party, called blind data miner. Each share does not reveal any information about the secret value, and therefore, a blind data miner does not learn the secret value. After individuals secret sharing to the blind data miners, the blind data miners run secure protocols to compute statistical functions on the data without revealing any information apart from the aggregated results.

The researchers performed the experiments on a real questionnaire dataset of 3158 respondents collected for a medical study and a simulated questionnaire dataset of 50 000 respondents, in the simulated local area network with three blind data miners. The results demonstrated the tool’s scalability.

Further the discussion with the audience about practical issues for privacy preserving while using data from distributed sources took place.

### 2.2 Day 2. Discussion between NSE and SSB on further collaboration

On the second day, NSE had a meeting with SSB to discuss potential use of privacy-preserving methods for further cooperation between the institutions. From SSB the following participants were presented in the discussion:

- **Anders Holmberg**, Head of Section for methodology. Confidentiality and programming meets methodology in this section. The privacy preserving has been placed at this Section
- **Øyvind Langsrud**, Researcher in Section for methodology
- **Diana-Cristina Iancu**, New employee in the Section for methodology
- **Jon Folkedal**, Enterprise architect in Section for architecture and IT-methods. Working mainly in Innovation project; a key person in the Modernization project
- **Johan Heldal**, Researcher in Section for methodology; involved in projects: disclosure control, RAID system, Kostra.

SSB collects large amounts of information about population. As a part of modernization work, SSB wants to reuse emerging data sources, such as transport and road sensor data, GPS, smart meters, and transaction data.
In the meeting, we mainly discussed the reuse of transaction data: combining it with data from registries makes it potentially sensitive.

SSB would like to compute count- and summation-based statistics stratified by demographic information, such as age / gender, and information about education from NUDB. Privacy is the main challenge for reuse of transaction data. The computation should not reveal any information apart from the results. In addition, the data custodians would like to maintain the access control over the data: the information providers do not want to share the information they provide with other providers while SSB should be able to produce statistics.

Important consideration for reuse of transaction data is that the supermarkets do not know demographic information and direct identifiers of their customers, unless the customer is a registered member. Analyzing only the data of customers with a membership might lead to bias. Therefore, SSB would like to link transactions at supermarkets with the bankcards transactions to be able to identify the personal numbers of the customers. Then, the personal numbers will be used to link the transaction data at the stores with the demographic information stored at SSB.

Privacy-preserving distributed statistical computation is considered as a potential solution. We started a discussion on applying this method for solving the problems mentioned above and agreed to continue the discussions on next coming sessions.

Additionally, the following ideas for further collaboration between NSE and SSB had been discussed:

- Reproduce SEDA statistics (Sentrale data fra allmennlegetjenesten), but without the collection of microdata. Last done by SSB in 2007.
- Paper (proposal): Vertically partitioned data (registry-based data) (FinaTek)
- New research-based center (SFI) for:
  - More efficient data gathering
  - Data from individuals (online)
  - Data from surveys
- Proposal to Horizon 2020
- Articles on disclosure control
- Experimental statistics in collaboration with Ministry of Health about health data and nutrition information from transactions data

NSE and SSB will continue the dialog on further collaboration.
3 Discussion

Collecting personally identifiable health information for health research can benefit both society and individuals via facilitating access to more effective health services [13]. The quality of statistical results is closely connected to amount and quality of available data. The data from one institution may not give sufficient statistical power or may not be diverse enough to address population and geographical heterogeneity [11]. Therefore, the data required, for example, for epidemiological and health services research are often distributed across multiple institutions. The distribution of data between several institutions prevents also the creation of huge centralized databases with extensive information on a single person. However, every such combination of data sources is a privacy risk for individuals.

To protect the identities and still be able to connect the data corresponding to the same individuals, pseudonymization can be used. Pseudonymization enables to uncouple specific data aspects from a data subject whereby the most identifying and/or sensitive data fields in the record are replaced by pseudonyms [14]. The distinction between anonymized and pseudonymized data is that pseudonymous data still can be re-identified (even indirectly and remotely), while anonymous data cannot be re-identified [15]. Cédric Nédélec, Data Protection Officer at PwC said: "A pseudonymisation technique which generates a secret key that is long and difficult to memorize (a combination of random characters) to which you apply what is known as a one-way function to the data (for instance, a cryptographic hash algorithm such as HMAC) will be more effective than a simple secret key cryptographic system" [14]. This solution offers some protection against curious data analysts. However, it is not sufficient to resist more determined and targeted attacks. Since the data fields of the records are not encrypted, it is possible to breach the privacy by comparing these fields to other datasets, for example, available data on the person’s gender, age, education, etc. [16]. Therefore, it is practically impossible to give any kind of security guarantee to a pseudonymization-based solution.

Another solution can be de-identification, i.e. removing the individual’s identifiers (such as names, addresses, identification numbers, etc.). The main challenge of de-identification is making a balance between re-identification risk and data utility. It is especially challenging in the context of distributed data. In this case, each institution locally de-identifies its data before sharing. However, the union of the de-identified data does not give the same result as working with the de-identified centrally collected data [11].

Herewith, even if the databases are pseudonymized/de-identified, records become larger in the joined databases. The chances to restore the identity are even higher if attackers connect additional information about the person from, for example, social networks.

For preserving privacy in statistical studies that analyze the data combined from several sources, the solution can be in combining secure multi-party computation (SMC) techniques with federated database systems. A federated database system is composed of several databases, which appear to function as a single entity. While processing a request, the system finds the component database which contains the data being requested and passes the request to it [17]. The goal of SMC is to perform all the required computations without revealing any microdata to the computing entities and mitigate the privacy risk [18], i.e. the private inputs of the input parties remain hidden from the computing parties and the result parties.

Secret sharing is a concept of hiding a secret value by splitting it into random parts and distributing these parts, called shares, to different parties, so that each party has only one share [19]. Each share (with no information about the secret value, i.e. without revealing any information on the individual level) is sent to a third party, called blind data miner. Secret sharing provides a way for one party to spread information on a secret such that all parties together hold full information, yet no single party has all the information [20]. Depending on the used secret sharing scheme, all or a known threshold
of shares are needed to reconstruct the original secret value [11]. Secure multi-party computation allows computing functions of secret-shared values so that each party learns only the corresponding function output value and no inputs of other parties. The blind data miners run secure protocols that compute statistical functions on the data producing the aggregated results. One benefit of this approach is that the data are not encrypted, and consequently do not run the risk of future breaches of encryption algorithms. This is in contrast to the solutions based on homomorphic encryption techniques. Also, by increasing the number of blind data miners, the likelihood of information security breaches is reduced.

With SMC, two personalized databases can be linked to aggregate the individual records into demographic groups, so that the resulting groups can be published. This is especially useful when the source databases are confidential and should not be openly linked.

In contrast to de-identification, with secure multi-party computation, both individuals’ and health institutions’ privacy can be protected without modifying or removing data variables. As a result, the quality of research data is not affected. Moreover, privacy-preserving solutions can potentially reduce the costs of conducting statistical studies on distributed data [21].
References

1. 18th WMA General Assembly, WMA DECLARATION OF HELSINKI – ETHICAL PRINCIPLES FOR MEDICAL RESEARCH INVOLVING HUMAN SUBJECTS. 1964: Helsinki, Finland.
10. Ursin, G.e.a., Protecting Privacy in Large Datasets—First We Assess the Risk, Then We Fuzzy the Data. Cancer Epidemiology, Biomarkers & Prevention, 2017.