Emnet: a system for privacy-preserving statistical computing on distributed health data

Meskerem Asfaw Hailemichael[a], Kassaye Yitbarek Yigzaw[a], Johan Gustav Bellika[b,c]

[a]*Department of Computer Science, UiT The Arctic University of Norway, Norway*
[b]*Department of Clinical Medicine, UiT The Arctic University of Norway, Norway*
[c]*Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, Norway*

Abstract

Reuse of health data for epidemiological and health services research have enormous benefits for individuals and society. However, patients' and health institutions' have privacy concerns. Yet, the commonly used de-identification and consent-based privacy-preserving methods have limitations.

In this paper we described three generic requirements for privacy-preserving statistical computing on distributed health data. Then, we described building blocks for implementation on horizontally partitioned data.

For each research project, a set of participant health institutions locally store data extracts for the researchers' criteria. The data across the institutions collectively make the project data, which we refer to as virtual dataset.

We decomposed count, mean, standard deviation, variance, covariance, and Pearson's r into summation forms and described as an abstract computation graph, where sub-computations are nodes. Generic APIs that can be invoked at runtime to execute a node against a virtual dataset are defined. Then we described a proof of concept implementation called Emnet.

Emnet demonstrates that horizontally partitioned data reuse can be possible while preserving patients' and institutions' privacy. More statistical analyses can easily be included into Emnet as far as they can be decomposed into summation forms.

Keywords:

Computation Graph, Data Reuse, EHR, Health Information System, Health Services Research, Privacy, Secondary Use, Statistical Computing, Secure Multi-party Computation, Secure Summation, Virtual Dataset

Introduction

The increasing use of electronic health record systems led to collection of a large amount of electronic health data at health institutions. In Norway, electronic health record (EHR) was first introduced in the late 1970s and now the usage has expanded to all GPs [1,2]. Reuse of health data collected for patient treatments have a huge potential for individuals and society through epidemiological and health services research including comparative effectiveness research, population-based surveillance, treatment safety, quality assurance [3,4]

However, misuse of data released for research could harm individuals and health institutions. Therefore, the privacy concerns remain to be the main challenges that have limited wide reuse of health data. Several jurisdictional, national and international ethical and legal regulations [5–8] have been passed to protect individual's privacy while enabling data reuse for research. In general, most regulations including the Norwegian Health Research Act [9] allow reuse of personal identifying data through informed consent and de-identified data without consent. In addition, a research ethics committee (e.g. REK in Norway) could allow reuse of personal identifying data without consent under certain conditions.

Informed consent could result in data bias due to demographic differences between consenters and non-consenters [5–8]. In addition, the time and cost requirements are often not feasible for large studies [10]. Data de-identification is a very important method for privacy protection. However, it is often challenged between minimizing probability of re-identification and increasing data utility [11]. In addition, these techniques do not protect the privacy of the health institutions, which is also considered a factor that limits data reuse [12,13].

The data available in one institution may not give sufficient statistical power, especially for rare diseases where there are only few cases at individual institution. In addition, it may not be diverse enough to address population heterogeneity. Population-based surveillances require data from multiple institutions that cover broad geographical area. Therefore, the data required for epidemiological and health services research is often distributed across multiple institutions.

Secure multi-party computation (SMC) techniques deals with the problem of a set of health institutions $H = \{H_1, H_2, \ldots, H_m\}$ who wish to jointly compute on their private data, while ensuring security properties, such as data privacy and correctness of output. These techniques only reveal computation results at the end of a computation [14]. As a result, both individuals' and health institutions' privacy can be protected.

Various statistical query tools and distributed research networks such as SAFTINET [15], EHR4CR [16], SHRINE [17], PopMedNet [18], and SCANNER [19] have implemented statistical analyses on data distributed across multiple health institutions. These tools, except SCANNER, only support statistical count. In addition, they do not protect the privacy of the health institutions, as individual institution level count is disclosed. In contrast, SCANNER supports more statistical analyses and has implemented computation techniques that release aggregated statistics of multiple institutions' data, which also protects individual institutions privacy.

In this paper we described a framework for privacy-preserving computing on distributed health data using SMC techniques, and its implementation called *Emnet*. *Emnet* enables statistical

analyses on data horizontally partitioned across multiple health institutions' EHRs. Currently, commonly used statistical analyses are implemented including *count, mean, standard deviation, variance, covariance,* and *Pearson's r*. However, the framework enables to easily add statistical analyses that can be decomposed into summation forms.

The remainder of this paper is organized as follows. Materials and Methods section describes the privacy requirements, building blocks of *Emnet*, and the Result section describes the design and implementation of Emnet and an experiment performed. The Discussion section discusses the main results of the implementation, and strength and limitation of the work presented in the paper.

Materials and Methods

In this section we have described the privacy requirements, and building blocks of the framework, which is divided into data preparation and statistical analyses.

Requirements for privacy preserving computing

We have formulated three requirements for privacy-preserving statistical analyses on data distributed across multiple institutions:

1. Any entity should not learn a combined statistics of < k number of institutions data. To protect the privacy of both individuals and health institutions, a computation should not reveal individuals' information and statistics on a single institution's data. Therefore, information revealed during a computation contains aggregate of individuals' data from $\geq k$ number of health institutions. The value of k depends on the privacy requirements of the health institutions.

2. *Semi-honest trust model.* Health institutions can be trusted to follow SMC protocols with their true data. However, no institution should be able to learn private information about individuals and health institutions from the messages exchanged during a computation.

3. *Must not depend on trusted third party.* No third party should be trusted to collect personal identifying sensitive data from health institutions. However, semi-trusted third party (STTP) could be used in a computation to improve computation efficiency and coordination. The STTP is only trusted not to collude with health institutions and follows SMC protocols. The STTP role can be given to the Norwegian Institute of Public Health or any other public authority.

Virtual dataset

As specified in the above requirements, a tool cannot use a trusted third party that collects the data required for a given research project. Therefore, each institution executes a project data query that contains inclusion and exclusion criteria, and the required data extracts. Then, data extracts are locally stored in a separate database. The data sets at all institutions collectively make the data required for the research project. A unique `project_id` is assigned to virtual datasets to correctly identify during analyses. Since the data are not stored in a central repository, we refer to these data sets as virtual dataset.

The focus of this paper is on horizontally partitioned data, therefore, each institution independently execute data query. However, for vertically partitioned data, virtual dataset creation requires record linkage techniques [20] to identify eligible patients and extract required data sets. Even when the data are horizontally partitioned, patients at the health institutions might not be mutually exclusive, especially when the health institutions are in geographically close area. For example, in Norway, residents can change their GP twice a year. As a result, an individual's data could be available at multiple GPs. Thus, virtual dataset creation on horizontally partitioned data also might require record linkage in order to identify and remove duplicate records. Duplicate detection is outside the scope of this paper.

OpenEHR is open standard specifications for EHR that enable to attain semantic interoperability. DIPS ASA[1], an EHR vendor that covers 70% of Norwegian hospital EHR market, is implementing openEHR based EHR. Norwegian ICT also deployed a Clinical Knowledge Manager (CKM)[2] registry for archetypes management and governance. Therefore, we assume that there is a drive towards wide use of openEHR archetype based EHRs in the health institutions.

Archetype Query Language (AQL) is the language developed to perform queries on openEHR based EHRs. It is neutral to specific implementation of EHRs, as far as the EHRs are based on openEHR specifications. Therefore, following our assumption of openEHR based EHRs across the health institutions, we have used AQL as a language to specify research projects' data query in the virtual dataset creation.

In this paper, we have implemented *Emnet* using an openEHR repository called Think!EHR[3]. Think!EHR is Java implementation of the latest openEHR specification. We persist openEHR compatible EHR extracts into the platform and execute queries specified using AQL.

Secure summation protocol

Yao introduced SMC in 1982 and since then it has been widely studied [21]. However, until the last decade practical implementation has been missing due to lack of efficient protocols. Specialized protocols (i.e. secure summation [22–24], and secure scalar product [25]) are designed to achieve better efficiency by utilizing specific properties of a computation. Protocols using generic techniques (i.e. garbled circuit [26], Homomorphic encryption [27,28], and secret sharing [29]) are also improving. Therefore, practical implementation of SMC tools are starting to appear [30].

SMC protocols are designed to provide security guarantee against a specific adversarial model (i.e. semi-honest, covert, or malicious adversary). Complex techniques are used to ensure stronger security guarantee (covert and malicious adversary), often at the cost of computation efficiency [31]. Therefore, protocols secure against semi-honest adversary are more efficient and scalable, and sufficient for joint computation between health institutions, as we assume health institutions can be trusted to follow an SMC protocol.

Secure summation is one of the most commonly studied protocol and a building block for several secure computations [32]. Secure summation protocols are designed using different techniques, such as secrete sharing [33,34], Homomorphic encryption [35], and adding random number on a private value. Two random number based protocols that are secure against semi-honest adversary are presented below.

[1] https://www.dips.no/
[2] http://arketyper.no/ckm/
[3] http://www.marand-think.com/

Simple Secure sum

Simple secure sum protocols are implemented based on adding random number on a private value before sending to another institution [24,36]. A coordinator sends a random number R to the first node. The first node adds its private value S_1 on R and passes the result $R + S_1$ to the second node. The second node does the same and passes the result $R + S_1 + S_2$ to the third node. Finally, the coordinator subtracts R from the value received from the last node $R + S_1 + S_2 + \ldots + S_n$ to find the true sum of the private values $S_1 + S_2 + \ldots + S_n$.

The protocol is efficient because it: (1) uses a simple technique; (2) only require equal number of communication as the number of nodes; and (3) has linear increase in number of communications with increase in number of nodes. However, the protocol does not ensure privacy, if node i and $i + 2$ collude to find a private value of node $i + 1$. Ensuring privacy against colluding nodes is a common challenge [33].

SINE (Secured Intermediate iNformation Exchange)

Shuwang et al. [37] implemented a random number based protocol with a better collusion resistance. A coordinator sends a random number R_c to the first node. The first node adds its own random number R_1 on R_c and passes the result $(R_c + R_1)$ to the second node. The second node does the same and sends the result $(R_c + R_1 + R_2)$ to the third node. Finally, the coordinator subtracts R_c from the value received from the last node $(R_c + R_1 + R_2 + \cdots + R_n)$ to find the sum of the random numbers $(R_1 + R_2 + \cdots + R_n)$. Subsequently, all nodes send the sum of their private value and their random number $(R_1 + S_1, R_2 + S_2, \ldots,$ and $R_n + S_n)$ to the coordinator. To find the sum of private values, the coordinator sums together these values and subtracts the sum of random numbers $((R_1 + R_2 + \cdots + R_n) + (S_1 + S_2 + S_n)) - (R_1 + R_2 + \cdots + R_n)$.

The SINE protocol only reveals the sum of all institutions' private value. The protocol provides security guarantee even when $n - 1$ nodes, other than the coordinator, collude with each other. However, if a coordinator collude with node $i + 1$, it is possible to learn private information of institution i. For example, if the coordinator receives $R_c + R_1$ from node 2, it is possible to calculate R_1 and consequently calculate S_1 from $R_1 + S_1$ received from node 1. Therefore, unless the coordinator colludes with other nodes, the protocol remains secure. The protocol only trusts that the coordinator follows the protocol and don't collude with other nodes.

In the simple secure sum protocol, collusion of any two (i and $i + 2$) nodes enables to learn private information of node $i + 1$. However, in the SINE protocol, the collusion should be between the coordinator and another node. We argue that it is easier to keep one node (the coordinator) secure from outside adversary, therefore, the protocol have stronger security guarantee. This is achieved at the cost of increased number of communications $(2n - 1)$ and arithmetic additions. In general, choice of a secure protocol requires a balance between the required security guarantee and computation efficiency.

As a result, in this paper, we chose the SINE protocol for our implementation of privacy preserving distributed computing tool as it satisfies our requirements described above. And the coordinator in the protocol is designated as STTP in the requirements.

Computation graph

A large number of linear and non-linear statistical analyses can be decomposed into sub-computations of summation form [38]. Therefore, each sub-computation can be computed with subset of the available data and the results can be sum together to find the overall result. This makes sub-computations suitable to be parallelized [39–41].

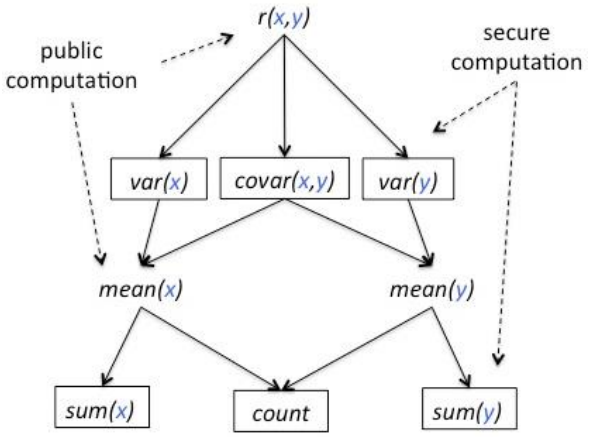
In this subsection, we have described decomposition of *count*, *mean*, *variance*, *standard deviation*, *covariance*, and *Pearson's r*, and how the decomposed statistical analyses can be computed in a privacy-preserving manner.

Let us assume three health institutions $\{H_1, H_2, H_3\}$ have horizontally partitioned data where each health institution has data of a unique set of patients that satisfied an inclusion and exclusion criteria. Let us further assume that the patients' ids at each institution are in the range of $[1, i]$, $[i + 1, n]$, and $[n + 1, m]$ (where $i > 0, n > i$ and $m > n$) respectively. The values of variables x and y are required for analyses.

Abstract computation graph

The statistical analyses chosen in this paper depend on one another: (1) *mean* x (y) depends on *sum of* x (y) and *count*; (2) *variance* of x (y) depends on *mean* of x (y); (3) *covariance* of x and y depends on *mean* of x and y; and (4) Pearson's r of x and y depends on *covariance* and *variance* of x and y. These dependencies are described as abstract computation graph shown in Figure 1. In the computation graph, the nodes represent statistical computations and the edges point to the direction of dependency between nodes. The dependency indicates that a node can be computed after computation of all the lower nodes that it depends on. For example, variance can be computed after summation and count.

As shown in Figure 1, there are two types of computations, such as secure and public computations. Computations that are in a box should be securely computed on individuals' data; and computations outside a box can be computed anywhere since they are based on only lower level nodes' results. Note that computation results of the nodes are considered as non-sensitive information.



The arrows point to the direction of dependency.

Figure 1 – Computation graph of summation, count, variance, covar-iance, and Pearson's r

The abstract computation graph does not have concrete information, such as where the input data are, and how the computation on each node is executed. How each analysis can be securely computed is described in the following subsection.

Table 1 – The operations provided by APIs that are implemented by different components of Emnet

Operations and parameters	Description
$LocalCompute\ (project\_id, equation,$ input$_values, variables, result_id)$	Locally executes an equation on individual patients' data
$SecureSum(project\_id,$ protocol$, addresses, result_id)$	Jointly run secure summation protocol on the results of $LocalCompute()$
$PublicCompute(project\_id, equation,$ input$_values, variables)$	Locally executes an equation on results of lower branch nodes on the graph

Concrete computation graph

Summation is the smallest statistical analysis; and other statistical analyses will be developed based on it. Summation of patients' values of x is shown in equation 1a. It can be expressed as equation 1b, where each institution locally sums their patients' values of x_j and then the local summation results of all institutions will be added together to find the total sum. The summation result from individual institution contains aggregate of its patients' data. Therefore, releasing it will not risk individuals' privacy. However, it can be considered private information of the health institution. Institutions privacy concerns can be avoided by using secure summation techniques that enable joint summation between institutions on their local summation results and only reveal the total summation result.

$$\boldsymbol{sum}(x) = \sum x_j \tag{1a}$$

$$\boldsymbol{sum}(x) = \sum_{j=1}^{i} x_j + \sum_{j=i+1}^{n} x_j + \sum_{j=n+1}^{m} x_j \tag{1b}$$

Therefore, the summation is divided into local computation and secure joint computation. Summation of individual level value is moved to where the data are located and can be computed much more efficiently. In contrast, since secure computations are more resource demanding, they are only used to aggregate local summation results.

A total count of eligible patients is a secure computation that is calculated from the sum of eligible patients in each institution. As shown in equation 2, each institution counts their local patients and then the local counts from all institutions are summed together using secure summation protocol.

$$\boldsymbol{count} = \boldsymbol{count}(H_1) + \boldsymbol{count}(H_2) + \boldsymbol{count}(H_3) \tag{2}$$

As shown in equation 3, *mean* of x is a public computation that is calculated from $sum(x)$ and $count$ results.

$$\boldsymbol{mean}\ (x) = \frac{\boldsymbol{sum}(x)}{\boldsymbol{count}} \tag{3}$$

Variance of x is a secure computation that is calculated from individuals' value of x_j and *mean* of x, as shown in equation 4a. Variance of an individual patient's value x_k is expressed in equation 4b. Variance can be expressed in equation 4c by substituting equation 4b into 4a, which becomes a summation problem and can be calculated in the same manner as the summation in equation 1a.

$$\boldsymbol{var}(x) = \frac{1}{count}\sum(x_j - \boldsymbol{mean}\ (x))^2 \tag{4a}$$

$$\boldsymbol{var}(x_k) = \frac{1}{count}(x_k - \boldsymbol{mean}\ (x))^2 \tag{4b}$$

$$\boldsymbol{var}(x) = \sum \boldsymbol{var}(x_j) \tag{4c}$$

A shown in equation 4d, standard deviation of x is a public computation that is calculated from variance of x result.

$$\boldsymbol{sdv}(x) = \sqrt{var(x)} \tag{4d}$$

Covariance of x and y is a secure computation that is calculated from individuals' value of x_j and y_j, and *mean* of x and y, as shown in equation 5a. Covariance of an individual patient's values of x_k and y_k is expressed in equation 5b. Covariance can be expressed in equation 5c by substituting equation 5b into 5a, which becomes a summation problem and can be calculated the same as the summation in equation 1a.

$$\boldsymbol{covar}(x,y) = \frac{1}{count}\sum(x_j - \boldsymbol{mean}\ (x))(y_j - \boldsymbol{mean}\ (y)) \tag{5a}$$

$$\boldsymbol{covar}(x_k,y_k) = \frac{1}{count}(x_k - \boldsymbol{mean}\ (x))(y_k - \boldsymbol{mean}\ (y)) \tag{5b}$$

$$\boldsymbol{covar}(x,y) = \sum \boldsymbol{covar}(x_j,y_j) \tag{5c}$$

Pearson´s r of x and y is a public computation that is calculated using covariance and variance values, as shown in equation 6a. Substituting covariance and variance equations (4a and 5a) into equation 6a, Pearson's r will be simplified to equation 6b.

$$\boldsymbol{r}(x,y) = \frac{\sum(x_j - \boldsymbol{mean}(x))(y_j - \boldsymbol{mean}(y))}{\sqrt{\sum(x_j - \boldsymbol{mean}(x))^2 \sum(y_j - \boldsymbol{mean}(y))^2}} \tag{6a}$$

$$\boldsymbol{r}(x,y) = \frac{\boldsymbol{covar}(x,y)}{\sqrt{\boldsymbol{var}(x)\boldsymbol{var}(y)}} \tag{6b}$$

The abstract computation graph shown in Figure 1 has a high level of abstraction and does not contain concrete computation details. It should be mapped to concrete computation graph for privacy preserving computing on a distributed data. Based on the discussions above, we have defined generic Application Programming Interfaces (APIs) for mapping from abstract to concrete computation graph at runtime. The APIs support the operations shown in Table 1.

As we have discussed above, execution of the secure computations on the abstract computation graph contain local and joint secure computations. Therefore, we have defined an API, called SecureComp API. Secure API includes two operations, such as $LocalCompute()$ and $SecureSum()$. Each secure computation node on the graph is mapped to consecutive execution of these operations. Since the secure computations are on private data, the API will be implemented at the health institutions.

1. $LocalCompute(project\_id, equation,$ input$_values, variables, result_id)$:

`project_id` is the id for a project that identify the virtual dataset on which the computation run; `equation` is name of the equation to be computed; `variables` are names of the variables to be computed on; $input\_values$ are results of lower level statistical analyses on the graph that the `equation` depends; and `result_id` is a unique id that will be assigned to the execution result. For example, to calculate variance of x, the `equation` is $var(x)$, $input\_values$ is value of *mean* of x, `variables` is x, and the `result_id` is a unique id.

2. $SecureSum(project\_id,$ protocol$, addresses, result_id)$:

`project_id` is the id for a project that identify the virtual dataset on which the computation run; `protocol` is name of a secure computation protocol to be used; `addresses` are addresses of peer health institutions that jointly compute the protocol; and `result_id` is a unique id for $LocalCompute()$ results that are jointly sum together. For example, to calculate variance of x, `addresses` are lists of addresses of $\{H_1, H_2, H_3\}$, `protocol`

is SINE secure summation protocol, and `result_id` is the same id assigned during execution of `LocalCompute()`.

As we have discussed above, execution of the public computations on the abstract computation graph are computed using only lower level nodes' computation results as input, that are not sensitive. Therefore, these computation can be computed either at the health institutions, STTP or client application, where the inputs are available. Therefore, we have defined an API called PublicComp API that includes `PublicCompute()` operation.

3. `PublicCompute(project_id, equation,` input`_values, variables)`:
`equation` is name of the statistical analysis to be computed; `input_values` are results of lower level statistical analyses on the graph that the `equation` depends; `project_id` is an id that enable to identify the `input_values` of a project; `variables` are names of the variables to be computed on. For example, to calculate *mean* of x, `equation` is *mean*, `variables` contain x and `input_values` are $sum(x)$ and *count* values.

Results

Design and Implementation

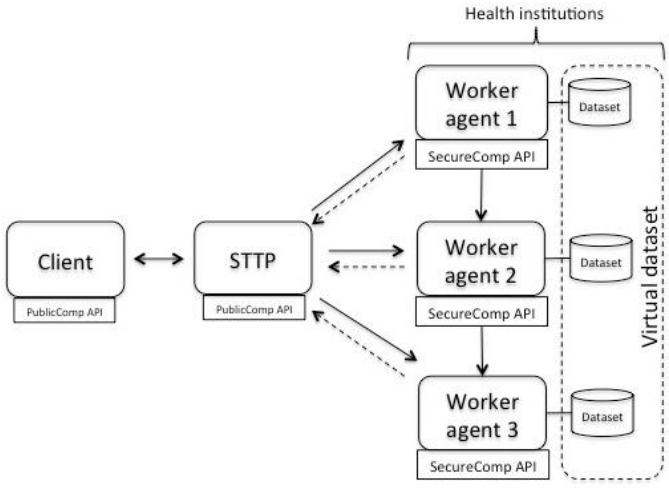
This section describes the design and proof of concept implementation of *Emnet* using the framework described above. Figure 2 shows main components of the tool.

Client – The Client is a web client application interface into *Emnet* and enables to specify a research project's data query, and statistical function and variables. It implemented the abstract computation graph and the Public API.

STTP – The STTP is a Java application gateway between the Client application and the health institutions, and it coordinates the overall executions. It implemented abstract computation graph, secure summation protocol and the Public API.

Worker agent – The Worker agent is a Java application that will be deployed at each health institution. It implemented secure summation protocols and the Secure API.

Emnet supports data preparation (*Virtual dataset creation*) and statistical analyses phases that are often required by research projects.



Client = Web application
STTP = Semi-Trusted Third Party

Figure 2 – Overall architecture of Emnet

Virtual dataset creation – a researcher specifies a data query on the Client application using the interface shown in Figure 3. The Client transforms into AQL and submits the query to the STTP who broadcasts it to each Worker agent. The Worker agents execute the AQL query against local openEHR and store the results locally in a MySQL database. Then, Worker agents reply the status of the query to STTP. The STTP executes descriptive statistics (currently only count of eligible patients)on the virtual dataset and returns results to the Client application.

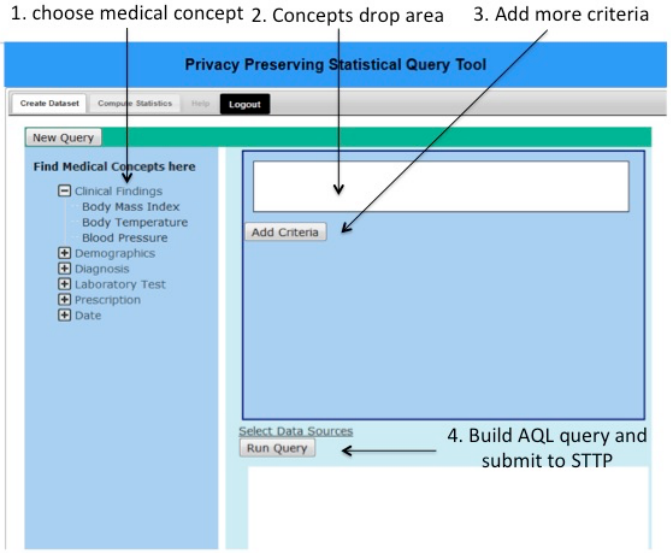


Figure 3 – Client interface to specify and execute virtual dataset cre-ation

Statistical analyses – similar to traditional statistical analyses tools, such as R[4] and SPSS[5], the user can specify the statistical function and variables on the Client application using the interface shown in Figure 4. If the requested statistical function is a public computation, for example *mean* of x, and if the lower branch of the abstract computation graph, such as count and $sum(x)$ are already calculated, the Client calls the public computation API. Otherwise the Client application submits the request to the STTP. STTP maps the required nodes on the abstract computation graph into concrete computation, by calling either the local Public API or Secure API at the Worker agents. The Worker agents execute the API calls on the local database. Finally, STTP returns the results to the Client.

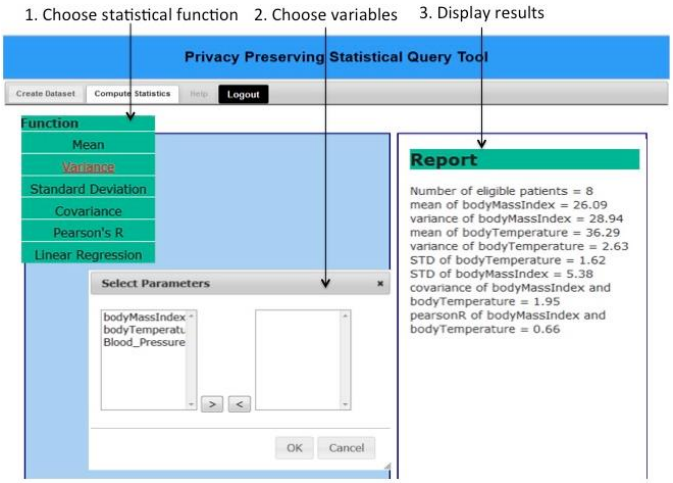


Figure 4 – Client interface to specify and execute statistical function on a virtual dataset

[4] http://www.r-project.org/

[5] http://www.spss.co.in/

Communication technology

In this section the technology used for communication between the different components of the architecture is described. *Emnet* is part of the Snow project[6], which is a distributed health data processing infrastructure deployed at multiple health institutions and labs in Norway [42]. The system implemented message-oriented communication using the Extensible Messaging and Presence Protocol (XMPP) [43]. The choice of the XMPP is due to the following reasons.

All healthcare service providers (i.e. GPs and hospitals) in Norway are connected via Norwegian Health Network, which is aimed to enable secure electronic communication between health institutions[7] The local networks of health institutions are considered more secure. Therefore, an institution should initiate all communications requests. The Snow system (40) has several software agents running at the health institution. Thus, each agents need to have its own address to receive requests sent to it.

XMPP technology is based on client/server architecture, similar to the SMTP protocol, where clients are interconnected through relaying servers. Therefore, each component contains an XMPP client identified by Jabber Id (JID) for communication. Each client authenticates using signed certificate and connects to the server, and the connection lasts long. Therefore, a client has address and connections are initiated from the health providers. In addition, XMPP enables point-to-point (i.e. between STTP and Worker agents, between Worker agents, and between Client and STTP), and multi-user (i.e. STTP broadcasts to Worker agents) messaging. In this paper, XMPP clients of the Worker agent and STTP are implemented using Smack library and on the Client web application Strophe library is used. Openfire server is used as XMPP server.

XMPP is a simple protocol that communicates over TCP sockets using XML messages. In addition, we have designed an XML message protocol that defines virtual dataset and statistical analysis requests and responses. The XML messages are sent inside XMPP XML message stanza.

Experiment

An experiment has been done based on a use case scenario designed to compute the correlation between human body temperature and body mass index. First, the two necessary archetypes, *Body Mass Index* and *Body Temperature*, were selected from Norwegian CKM and a template containing these archetypes was designed. Then, we prepared test openEHR data sets using the template and a virtual environment that simulates the real working environment with three distributed EHRs. On this virtual environment, we computed *Mean, Variance, Standard Deviation, Covariance* and finally *Pearson's r* (correlation) of Body Mass Index and Body Temperature.

Discussion

We have described a generic framework and implementation of *Emnet* for computing on horizontally partitioned distributed health data. The developed framework satisfies the three privacy requirements we defined to preserve the privacy of both individuals and health institutions. In addition, it enables institutions to maintain strong control over who compute, what analyses, and on what data. However, access control is outside the scope of this paper.

Currently, *Emnet* implements *count, mean, standard deviation, variance, covariance,* and *Pearson's r.* It can easily be extended to include more statistical analyses as far as they can be decomposed into summation form.

The building blocks for the framework can be divided into data preparation and statistical analyses phases. For each research project, health institutions locally store data extracts for criteria specified by the researchers'. These data extracts across the institutions collectively make the project data, which we refer as virtual dataset. Since a common data model is required across the health institutions, we make an assumption that the health institutions have openEHR-based health record systems.

We decomposed the statistical equations into sub-computations of summation form and created dependencies between them. We expressed these dependencies as an abstract computation graph, where each node represents a sub-computation. In order to execute a statistical analysis against a virtual dataset, all the lower level nodes should be executed first. We have described how the nodes can be executed using simple arithmetic and/or secure summation protocol. Then, we created an abstraction using APIs that can be invoked at runtime to execute a node on the abstract computation graph.

Comparison of *Emnet*'s computation efficiency with traditional statistical analyses tools such as R and SPSS, where the data are centrally stored, is invaluable. Evaluation of the computation efficiency will be a future work. However, we hypothesis that Emnet is efficient, because (1) computations that require individual patients data are computed locally and all health institutions compute in parallel; and (2) only aggregations of local results are computed using simple secure summation protocol.

In general, as the number of participating health institutions increases, efficiency of statistical analyses might decrease. However, in [44] we have described a technique to maintain constant efficiency independent of the number of participants. As a result, *Emnet* can be scalable. Implementation of the technique into the tool will be a future work.

Despite the benefits of health data reuse, quality of data and their suitability for research is a concern [45]. The main benefit of the virtual dataset is that it enables to do either clerical review or run computer programs to improve the data quality without modifying the original data. In addition, it supports to store preprocessing and intermediate results of statistical analyses.

In contrast to de-identification [11], the technique presented in the paper preserves privacy without modifying or removing data variables. As a result, the quality of research data is not affected due to the privacy-preserving computation.

Both the public [46] and healthcare professionals [47,48] demonstrated positive view towards reuse of health data for research as long as the privacy and other concerns are addressed. Emnet could increases health institutions' and patients' willingness for reuse of their data for research. As a result, enormous benefits of health data reuse can be unlocked.

[6] The Snow system client application is available at http://snow.telemed.no/

[7] https://www.nhn.no/english/Pages/about.aspx

In general, Emnet will increase researchers' access to health data with the following added benefits, (1) better privacy; (2) quality of data; and (3) minimized time and cost to collect data. More health research enables to improve effectiveness, efficiency and quality of care. Consequently, the public will benefit.

The framework described in this paper can be applied for any domain outside health, where there is a need for joint computation on private data while maintaining privacy. In addition, it is light weighted for implementation on small devices, such as smart phones to jointly compute on apps' data of a set of individuals.

Acknowledgments

The second author was supported by the Center for Research-Based Innovation, Tromsø Telemedicine Laboratory, through Research Council of Norway Grant No. 174934. We would like to acknowledge the invaluable contribution of Luis Marco Ruiz on the use of openEHR. We are obliged to acknowledge Marand, Slovenia for letting us use their Think!EHR platform. We also would like to acknowledge the support from Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, and UiT The Arctic University of Norway.

References

[1] Dobrev A, Haesner M, Husing T, Korte BW, Meyer I. Benchmarking IT Use Among General Practitioners in Europe. Bonn,Germany: European Commission; 2008.

[2] Christensen T, Faxvaag A, L\a erum Hallvard, Grimsmo A. Norwegians GPs' use of electronic patient record systems. International Journal of Medical Informatics 2009;78:808–14.

[3] Selby JV, Krumholz HM, Kuntz RE, Collins FS. Network News: Powering Clinical Research. Sci Transl Med 2013;5:182fs13–182fs13. doi:10.1126/scitranslmed.3006298.

[4] Friedman CP, Wong AK, Blumenthal D. Achieving a Nationwide Learning Health System. Sci Transl Med 2010;2:57cm29–57cm29. doi:10.1126/scitranslmed.3001456.

[5] Tu JV, Willison DJ, Silver FL, Fang J, Richards JA, Laupacis A, et al. Impracticability of informed consent in the Registry of the Canadian Stroke Network. N Engl J Med 2004;350:1414–21. doi:10.1056/NEJMsa031697.

[6] Young AF, Dobson AJ, Byles JE. Health services research using linked records: who consents and what is the gain? Australian and New Zealand Journal of Public Health 2001;25:417–20. doi:10.1111/j.1467-842X.2001.tb00284.x.

[7] Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data Linkage: A powerful research tool with potential problems. BMC Health Services Research 2010;10:346. doi:10.1186/1472-6963-10-346.

[8] Carter K, Shaw C, Hayward M, Blakely T. Understanding the determinants of consent for linkage of administrative health data with a longitudinal survey. Kōtuitui: New Zealand Journal of Social Sciences Online 2010;5:53–60.

[9] Norwegian Ministry of Health. ACT 2008 - 06 - 20 no. 44: Act on medical and health research (the Health Research Act). vol. no. 44. 2009.

[10] Kho ME, Duffett M, Willison DJ, Cook DJ, Brouwers MC. Written informed consent and selection bias in observational studies using medical records: systematic review. BMJ 2009;338:b866–b866. doi:10.1136/bmj.b866.

[11] Wu FT. Defining Privacy and Utility in Data Sets. Rochester, NY: Social Science Research Network; 2012.

[12] El Emam K, Mercer J, Moreau K, Grava-Gubins I, Buckeridge D, Jonker E. Physician privacy concerns when disclosing patient data for public health purposes during a pandemic influenza outbreak. BMC Public Health 2011;11:454.

[13] El Emam K, HU J, Mercer J, Peyton L, Kantarcioglu M, Malin B, et al. A secure protocol for protecting the identity of providers when disclosing data for disease surveillance. Journal of the American Medical Informatics Association 2011;18:212–7. doi:10.1136/amiajnl-2011-000100.

[14] Lindell Y, Pinkas B. Secure multiparty computation for privacy-preserving data mining. Journal of Privacy and Confidentiality 2009;1:5.

[15] Lisa M. Schilling BMK. Scalable Architecture for Federated Translational Inquiries Network (SAFTINet) Technology Infrastructure for a Distributed Data Network. eGEMS 2013;1:Article 11. doi:10.13063/2327-9214.1027.

[16] Voets D. EHR4CR. Initial EHR4CR architecture and interoperability framework specifications. EHR4CR, 2012.

[17] McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies. PLoS ONE 2013;8:e55811. doi:10.1371/journal.pone.0055811.

[18] Vogel J, Brown JS, Land T, Platt R, Klompas M. MDPHnet: Secure, Distributed Sharing of Electronic Health Record Data for Public Health Surveillance, Evaluation, and Planning. Am J Public Health 2014;104:2265–70. doi:10.2105/AJPH.2014.302103.

[19] Kim KK, Browe DK, Logan HC, Holm R, Hack L, Ohno-Machado L. Data governance requirements for distributed clinical research networks: triangulating perspectives of diverse stakeholders. J Am Med Inform Assoc 2014;21:714–9. doi:10.1136/amiajnl-2013-002308.

[20] Christen P. Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer-Verlag Berlin Heidelberg; 2012.

[21] Yao AC. Protocols for secure computations. Proceedings of the 23rd Annual Symposium on Foundations of Computer Science, Washington, DC, USA: IEEE Computer Society; 1982, p. 160–4. doi:10.1109/SFCS.1982.88.

[22] Benaloh JC. Secret Sharing Homomorphisms: Keeping Shares of a Secret Secret (Extended Abstract). In: Odlyzko AM, editor. Advances in Cryptology — CRYPTO' 86, Springer Berlin Heidelberg; 1987, p. 251–60.

[23] Karr AF, Lin X, Sanil AP, Reiter JP. Secure Statistical Analysis of Distributed Databases. In: Wilson AG, Wilson GD, Olwell DH, editors. Statistical Methods in Counterterrorism, Springer New York; 2006, p. 237–61.

[24] Andersen A, Yigzaw KY, Karlsen R. Privacy preserving health data processing. 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom), 2014, p. 225–30. doi:10.1109/HealthCom.2014.7001845.

[25] Xu F, Zeng S, Luo S, Wang C, Xin Y, Guo Y. Research on Secure Scalar Product Protocol and Its' Application. 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), 2010, p. 1–4. doi:10.1109/WICOM.2010.5601452.

[26] Yao AC. Protocols for Secure Computations. Proceedings of the 23rd Annual Symposium on Foundations of Computer Science, Washington, DC, USA: IEEE Computer Society; 1982, p. 160–4. doi:10.1109/SFCS.1982.88.
[27] Gentry C. Fully Homomorphic Encryption Using Ideal Lattices. Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing, New York, NY, USA: ACM; 2009, p. 169–78. doi:10.1145/1536414.1536440.
[28] Paillier P. Public-Key Cryptosystems Based on Composite Degree Residuosity Classes. In: Stern J, editor. Advances in Cryptology — EUROCRYPT '99, Springer Berlin Heidelberg; 1999, p. 223–38.
[29] Chaum D, Crépeau C, Damgard I. Multiparty Unconditionally Secure Protocols. Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing, New York, NY, USA: ACM; 1988, p. 11–9. doi:10.1145/62212.62214.
[30] Bogdanov D. Sharemind: programmable secure computations with practical applications. Thesis. Tartu University, 2013.
[31] Yigzaw KY, Bellika JG. Evaluation of secure multi-party computation for reuse of distributed electronic health data. 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2014, p. 219–22. doi:10.1109/BHI.2014.6864343.
[32] Youwen Z, Liusheng H, Wei Y, Xing Y. Efficient Collusion-Resisting Secure Sum Protocol. Chinese Journal of Electronics 2011;20.
[33] Urabe S, Wong J, Kodama E, Takata T. A High Collusion-resistant Approach to Distributed Privacy-preserving Data Mining. Proceedings of the 25th Conference on Proceedings of the 25th IASTED International Multi-Conference: Parallel and Distributed Computing and Networks, Anaheim, CA, USA: ACTA Press; 2007, p. 326–31.
[34] Shepard S, Kresman R, Dunning L. Data Mining and Collusion Resistance. Proceedings of World Congress on Engineering 2009 2009;1.
[35] Drosatos G, Efraimidis PS. Privacy-Preserving Statistical Analysis on Ubiquitous Health Data. In: Furnell S, Lambrinoudakis C, Pernul G, editors. Trust, Privacy and Security in Digital Business, Springer Berlin Heidelberg; 2011, p. 24–36.
[36] Karr AF, Karr AF, Lin X, Lin X, Sanil AP, Sanil AP, et al. Secure Regression on Distributed Databases. J Computational and Graphical Statist 2004;14:263–79.
[37] Shuang Wang XJ. EXpectation Propagation LOgistic REgRession (EXPLORER): Distributed Privacy-Preserving Online Model Learning. Journal of Biomedical Informatics 2013. doi:10.1016/j.jbi.2013.03.008.
[38] Kearns M. Efficient Noise-Tolerant Learning From Statistical Queries. Journal of the ACM, ACM Press; 1993, p. 392–401.
[39] Chu C, Kim SK, Lin Y-A, Yu Y, Bradski G, Ng AY, et al. Map-reduce for machine learning on multicore. Advances in Neural Information Processing Systems 2007;19:281.
[40] Das S, Sismanis Y, Beyer KS, Gemulla R, Haas PJ, McPherson J. Ricardo: integrating R and Hadoop. Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, New York, NY, USA: ACM; 2010, p. 987–98. doi:10.1145/1807167.1807275.
[41] Duan Y. P4P: A Practical Framework for Privacy-Preserving Distributed Computation. PhD thesis. University of California, 2007.
[42] Bellika JG, Henriksen TS, Yigzaw KY. The Snow system - a decentralized medical data processing system. In: Llatas CF, García-Gómez JM, editors. Data Mining in Clinical Medicine, Springer; 2014.
[43] Saint-Andre P, Smith K, Tronçon R. XMPP: The Definitive Guide. O'Reilly Media, Inc.; 2009.
[44] Yigzaw KY, Bellika JG, Andersen A, Hartvigsen G, Fernandez-Llatas C. Towards Privacy-preserving Computing on Distributed Electronic Health Record Data. Proceedings of the 2013 Middleware Doctoral Symposium, New York, NY, USA: ACM; 2013, p. 4:1–4:6. doi:10.1145/2541534.2541593.
[45] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2012. doi:10.1136/amiajnl-2011-000681.
[46] Institute of Medicine (US) Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. Washington (DC): National Academies Press (US); 2009.
[47] Hopf YM, Bond C, Francis J, Haughney J, Helms PJ. Views of healthcare professionals to linkage of routinely collected healthcare data: a systematic literature review. J Am Med Inform Assoc 2014;21:e6–10. doi:10.1136/amiajnl-2012-001575.
[48] Hopf YM, Bond C, Francis J, Haughney J, Helms PJ. "The more you link, the more you risk …" – a focus group study exploring views about data linkage for pharmacovigilance. Br J Clin Pharmacol 2014;78:1143–50. doi:10.1111/bcp.12445.

Address for correspondence

Kassaye Yitbarek Yigzaw
e-mail: kassaye.y.yigzaw@uit.no