UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

Faculty of Science and Technology
Department of Computer Science

Emnet: A System for Privacy-preserving Statistical Computation on Distributed Health Data

—

Meskerem Asfaw Hailemichael

INF-3997 Master's Thesis in Telemedicine and E-health - May 2015

Preface

There is a growing interest to reuse the enormous health data being collected at health institutions. Reuse of these data provides vast opportunities for healthcare research. However, it is facing multi-dimensional challenges where privacy and interoperability are the main challenges. Studies have shown that considerable progress is being made to deal with these challenges and to facilitate a wide scale implementation of data reuse. Yet, existing solutions have limitations that need to be addressed.

This thesis is dedicated to advance the road towards society that enjoys the benefit of privacy and better healthcare services through research. We have developed a technique for privacy-preserving statistical computation on distributed EHRs. The bases for requirements were scientific literatures and potential users.

In general, trust is the corner stone of our society and in particular for doctor-patient relationships. Without trust patients hesitate to disclose their medically relevant information, and be medically examined. Even worse, they may not access health services. Health data reuse should not weaken this trust. Privacy-preserving health data reuse techniques should be coupled with building trust in the society. Therefore, to emphasize the value of trust we named our system Emnet (taken from the Amharic[1] word “እምነት”), which means trust.

This thesis is part of the Snow project, which is an on-going research project at Norwegian Centre for Telemedicine (NST), University Hospital of North Norway (UNN). The Snow project is mainly focused on development of techniques and tools for health data reuse. The project has deployed the Snow system in general practitioner (GP) offices, hospitals and laboratories for disease surveillance. The system has a web client that informs GPs and patients about how communicable diseases spread chronologically and geographically. This thesis also extends the work in the Snow system with privacy-preserving statistical computation capability.

The implementation of the techniques has a web application where the user specifies research criteria to select a data set from distributed EHRs; and perform statistical computations on the selected data set without transferring data outside the health institutions. *Emnet* uses openEHR based EHRs to maintain interoperability; and secure multi-party computation (SMC) protocol to jointly compute among the components of *Emnet*. In doing so, it ensures privacy of individuals as well as health institutions participated in the research.

[1] Amharic is the federal working language of Ethiopia

First and foremost I thank God, the almighty, for the good health and wellbeing that were necessary to complete this work.

I offer my sincerest gratitude to my supervisor professor Johan Gustav Bellika who showed me the road and helped me grow as a professional. I am extremely thankful and indebted to him, for his sincere guidance and encouragement.

My deepest gratitude goes to my co-supervisor Luis Marco Ruiz for all the support and tireless dedication to the thesis work. Apart from being a supervisor, he kindly devoted his time to help me learn openEHR. I am also grateful to Stein Olav Skrøvseth, my co-supervisor for the valuable comments on the thesis.

Furthermore, I would like to thank Gro Berntsen and Nils Kolstrup for being available for discussion during the requirements specifications. My earnest gratitude goes to Kassaye Yitbarek Yigzaw for his invaluable support and encouragements during the thesis work. He was always available for discussions regardless of his own busy schedule. Special thank goes to Torje Starbo Henriksen for willingly sharing his knowledge on the technical aspects of the thesis and commenting on the thesis report.

I am pleased to acknowledge Marand d.o.o., Slovenia for providing Think!EHR platform and kind cooperation. I am also grateful to all the scholars on whose work I have learned. I wish to express my sincere thanks to the Norwegian quota scheme program, for giving me the opportunity and the financial support during this MSc. study. I am also obliged to acknowledge all the technical supports from the Department of Computer Science, UiT and NST.

Last but not least, I would like to thank my family for the blessings, encouragements and moral supports through the phone calls.

Meskerem Asfaw Hailemichael
Tromsø, Norway
May 2015

Abstract

Motivation: Despite its enormous benefits, EHR data reuse is limited because of multi-dimensional challenges where privacy comes on the forefront. Recently various privacy-preserving statistical computation tools have emerged. However, they have limited privacy guarantee and use ad-hoc techniques for privacy-preserving computation of statistical functions.

Purpose: The purpose of this thesis is to develop a system that enables to compute a wide variety of statistical functions on distributed EHRs, while preserving the privacy of patients and health institutions.

Materials and Methods: Systematic literature review of privacy-preserving techniques for health data reuse was performed to understand the state-of-the-art. The result of the review and meetings with users were used as sources of requirements. Agile methodology was used for implementation of a prototype system called Emnet.

Emnet uses openEHR-based EHRs as common data model to achieve interoperability among health institutions. We have prepared test openEHR data sets and a virtual environment that simulates the real working environment for testing.

Result: We have developed and tested privacy-preserving techniques for research *data set preparation* and *statistical computation*. The research eligibility criteria and required attributes are expressed as a computable query using Archetype Query Language (AQL), and each health institution executes the query and locally stores the resulting data set. The data sets are physically distributed across the health institutions, yet they collectively make the research data set, which we call *Virtual Dataset*.

Statistical computations on the *Virtual Dataset* are performed using two main techniques, (1) decomposition of statistical functions into summation forms and described as a computation graph; and (2) secure summation protocols.

Conclusion: The developed techniques enable statistical computation on distributed health data, while preserving the privacy of patients and health institutions. Currently, *mean, variance, Standard Deviation, Covariance* and *Pearson's r* are implemented in Emnet. However, the techniques are generic to implement more statistical functions, as long as they can be decomposed into summation forms. The work presented in this thesis contributes for advancement of privacy-preserving health data reuse. It is also relevant to other domains where they have similar requirements as health care.

Keywords: Computation Graph, Data reuse, EHR, Health Information System, Health Research, Privacy, Statistical Computing, Secure Multi-party Computation, Secure Summation, Virtual Dataset

Table of Contents

List of Figures

List of Tables

Abbreviations

EHR- Electronic health record

HIPAA- United States, the health insurance portability and accountability act

XMPP- Extensible Messaging and Presence Protocol

API - Application programming Interface

GP - General practitioner

SMC - Secure Multi-party Computation

Chapter 1 Introduction

1.1. Motivation and Background

The wide use of health information system resulted in a large amount of data collected at each health institution. Even though the primary purpose of data collection is patient treatment, it also presents opportunities to conduct healthcare research like population-based surveillance, treatment safety, comparative effectiveness research, quality assurance and learning health systems (1–4).

Research like population-based surveillances requires data from several institutions that cover broad geographical area. Moreover, 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. Therefore, the data required for epidemiological and health services research is found across various distributed databases (1,5).

Compared to the traditional research methods, EHR data reuse has a potential to ease health research as it reduces the cost and time needed for data collection. Despite its enormous benefits, health data reuse has multi-dimensional challenges. As the primary purpose of the collected data is for patient treatment, reusing it for research involves several challenges including privacy and security issues; legal and ethical issues; cross-institutional contracting policies and regulations; heterogeneity of the various databases (EHRs); and the quality and comprehensiveness of the data (the data collected for patient treatment may not satisfy a research protocol level quality) (6–9). Among these challenges, privacy and interoperability issues come in the forefront (1,9).

Privacy issue in this case is the fear of data owners (patients and healthcare givers) that their information might be misused. The issue of privacy is a threat to patients to an extent that they self-medicate their illnesses, take traditional medications, lying about health condition, change doctors, ask their information not to be registered in EHR, be unwilling to participate in clinical trials and be reluctant to give consent to any public health research (10,11).

Centralized and distributed approaches have been used to store health data for research purpose (6). The centralized approach involves collection of data from the various health institutions to a large centralized database. In contrast, distributed approaches involve computing on the distributed data, without moving individuals' data from the health institutions. The centralized approach is considered to be simple and less complicated because all attributes of the data needed for the research are collected in a single database (12). While others argue that the distributed approach has many practical advantages. These include, autonomy to the health institutions regarding who uses what, reduces security and privacy concerns (6,13–15). As a result, it encourages patients and health institutions to participate in research.

Implementing the distributed approach involves different privacy-preserving computation techniques such as secure multi-party computation (SMC). SMC is cryptography based computing technique for multiple parties to jointly compute on their secret values and reveal only the computation results at the end of the computation.

Ever since the concept of health data reuse was introduced, various EHR query tools and distributed research networks have been proposed (e.g. WICER, SCOAP-CERTN, SPAN, RPDR, SPIN, SHRINE, PopMedNet, SCANNER) (14,16–18). However, the privacy level and the supported statistical functions of the existing tools have limited their wider use.

The aim of this thesis is to develop a privacy-preserving tool that enables statistical computations on horizontally partitioned, distributed health data. It is a continuation of authors' previous academic work reported in (19). Mainly, in this thesis, more focus is given to privacy and implementation of statistical functions.

1.2. Scope and Research Problems

The general objective of the thesis is to answer the following research question:

How can a privacy-preserving statistical computation tool that enables statistical computations on distributed electronic health records be developed?

The above question is divided into sub-questions in order to better understand the research question and define the scope of the thesis. The sub-questions are explained in detail in the following sections.

A. Research data preparation

Traditionally a data set that fulfills a research inclusion and exclusion criteria are collected and stored in a centralized database. However, in the distributed approach the data should remain at the health institutions. Therefore, the following two main questions should be answered:

Question 1. *How can the research inclusion and exclusion criteria be specified?*

Question 2. *How can a research data (a Virtual Dataset) be created based on research criteria without moving the data outside the health institutions?*

B. Statistical Computation

Traditionally, statistical analyses are performed on data sets that are stored in central database. Nonetheless, performing privacy-preserving statistical computation on distributed data sets is not straightforward. Hence, we need to create a suitable environment for any non-technical researcher to carry out different statistical computations regardless of the distribution of the data sources. Therefore, the following question arises:

Question 3. *How can statistical computations be performed on the distributed data sets in a privacy-preserving manner?*

1.3. Assumptions and Limitations

The following paragraphs describe the main assumptions made in this thesis.

1. The EHRs are interoperable

It is commonly accepted that lack of interoperability is another major barrier for EHR data reuse (9,20,21). Different healthcare institutions use different information systems to store their records that make interoperability difficult. Interoperability, in healthcare, is defined as the "ability of information systems to work together within and across organizational boundaries in order to advance the effective delivery of healthcare for individuals and communities" (22). Interoperability refers to both syntactic (having the same physical data structure) and semantic (having the same meaning for a single concept) interoperability.

To ensure interoperability, the current strategic plan of Norwegian Health authorities is encouraging EHR vendors to adopt openEHR (23). For example, DIPS ASA (24), which is the provider of more than 70% of hospital EHRs and 1,500 primary care institutions' communication solutions, is implementing an openEHR-based EHRs (25). openEHR is an open standard specification that enables to attain semantic interoperability among electronic health records (26). Therefore, in this thesis, we assume that interoperability among the EHRs can be achieved by using the openEHR specifications.

2. Data are horizontally partitioned.

It is possible that a single patient receives treatments from different kinds of health institutions (primary and secondary healthcare), which makes the patient's records vertically partitioned across health institutions. Besides, in some cases, data duplication occurs even in horizontally partitioned data. For example, in Norway a patient is allowed to change his/her general practitioner (GP) twice in a year. Consequently, duplicate record of a single patient can be found in two or more institutions. However, in this thesis, we assume that the data are horizontally partitioned and no duplicate data exist, i.e. a patient record is found only in a single health institution.

3. Sufficient data quality

In data reuse, the researcher has no role in the planning and data collection that might minimize the data quality for research. Lack of data quality (which embraces completeness, correctness, concordance, plausibility, and currency) (27) is another challenge for EHR data reuse. But in this thesis, we assume that the quality of the data is acceptable for research.

4. Health institutions follow computation protocols honestly

Health institutions that share data for a computation are considered to be secure and honestly involved for the best interest of the public. Accordingly, we assume that the health institutions share their correct data and follow computation protocols. Yet, they might be curious to know about other health institutions' information. This assumption is known as semi-honest (honest-curious) adversary model (see section 2.2.2.).

1.4. Significance and Contribution

This thesis aims at enabling statistical analysis on distributed EHRs while keeping privacy of data owners (patients and health institutions). Results of the literature review we have done together with the users' requirements helped us to understand the knowledge gap that exists in the area. Even though data exists in electronic format almost everywhere, EHR data reuse have been limited by the different challenges explained above, where privacy is the leading one. Hence, the focus of this thesis is mainly on privacy. The resulting prototype demonstrates how to create data sets for research analysis (while the data remains at the health institutions) and perform statistical computations on the data sets in a privacy-preserving manner. In this context, privacy embraces the privacy of not only patients but also healthcare institutions. Moreover, this thesis is built based on the openEHR specification, which implements a multi-level modeling framework to facilitate interoperability. In addition, the thesis also contributed to scientific publication; A paper entitled "***Privacy-preserving Statistical Query and Processing on Distributed openEHR data***" (*Meskerem Asfaw Hailemichael, Luis Marco Ruiz and Johan Gustav Bellika*) is accepted in Medical Informatics Europe Conference (MIE 2015)[2]. The paper mainly contains the high level description of the solution together with the architecture developed in the thesis (see Appendix D).

Therefore, this thesis can be a good start to develop further research in the emerging area of privacy-preserving health data reuse. Moreover, it presents design techniques on query processing against distributed openEHR based EHRs.

[2] http://www.mie2015.es/

1.5. Organization of the Thesis

The remainder of the thesis is organized into the following chapters:

Chapter 2 – Theoretical Background

This chapter describes the-state-of-the-art and literature review in the privacy-preserving EHR data reuse area. It also gives an overview of privacy in health data reuse, the existing SMC protocols and the technical frameworks used in this thesis.

Chapter 3 – Methods

This chapter is dedicated to show the research paradigm, and materials and methods used in this thesis.

Chapter 4 – Requirements Specification

This chapter describes the sources of the requirements, and details about the functional and non-functional requirements of the thesis. .

Chapter 5 – Design

This chapter contains the architectural design of the prototype system. It also describes the protocol design for the different statistical functions implemented in this thesis.

Chapter 6 – Implementation

This chapter describes the implementation details of the design presented in the chapter 5. It also explains the technologies and platforms used to develop the prototype.

Chapter 7 – Testing and Results

This chapter describes the process of test-data preparation. It also presents the testing procedure and the result obtained.

Chapter 8 – Discussion

This chapter discusses the concepts developed and implemented in this thesis. It also describes the solution developed in comparison with similar studies identified in the literature review. Limitations of the thesis are also stated in this chapter.

Chapter 9 – Conclusion and Future Work

This is the last chapter that contains the concluding remarks in relation with the research problems and it also suggests the possible future direction of the work for further development and research.

Chapter 2 Theoretical Background

This chapter aimed to describe the state-of-the-art of EHR data reuse and privacy-preserving statistical query processing. The first section describes the terms and concepts in this field. The next section presents the process and results of the literature review and the final section explains the major query tools and distributed research networks that have been developed or in the process of development.

2.1. Privacy

2.1.1. What is privacy?

The meaning of the word privacy is complex and varies with context. Malin et al. (11) define the word privacy as constituent of various constructs, such as anonymity, confidentiality, and solitude. Anonymity is the ability to hide one's identity; confidentiality is the ability to share information with a second party without the information being publicly revealed; and solitude is the right to be left alone. HIPPA (28) define privacy as it "pertains to the collection, storage, and use of personal information and addresses the question of who has access to personal information and under what conditions". With respect to health information, privacy issues could also be described as unforeseen interferences in the research participants' personal life (29). The more general definition of privacy is given by Alan F. Westin (30), which states that *"Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others."*

2.1.2. Why privacy?

The growth of information technology has increased the ability to collect and manipulate large amount of data. Concurrent with this growth, the necessity for information linked from different sources for better health decision resulted in high degree of privacy risk (31). Even if there is a difference from culture to culture, the legal and ethical concerns of EHR data reuse are similar between countries (32). Similarly, EuroREACH[3] showed that privacy and legal issues are the major barriers in using health data for research in most of the EU member countries including Austria, Estonia, England, France, Israel, and Luxembourg (33).

Most ethical and legal regulations allow data reuse through informed consent or data de-identification (see detailed discussion in subsection 2.2.1.) (34). Consent has been used widely in healthcare research. Norwegian Health Research Act (35) also dictates that patient

[3] « an international collaboration to improve access to and use of healthcare data and to enhance cross-country comparisons of health system performance.»

identifiable data can be reused through informed consent. However, studies (36–39) have shown that there are significant socio-economic and demographic differences between consenters and non-consenters that have a potential to create systematic bias on the research result. Moreover, the time and money required to collect consent is impractical for studies covering large population (40).

Several ethical and legal regulations allow reuse of patient identifiable data without consent, under limited conditions, for the interest of public health. Some of the regulations are: HIPAA (28), FEAM (Federation of European Academies of Medicine) (41) and Norwegian Act on medical and health research (35). On the other hand, patients claim the importance of consent. For example, studies in Sweden (42) and Finland (43) on patient data reuse addressed the following two issues; 1) whether patient consent is needed to use Biobank data for research purpose 2) if data consented for one research is used for another research. Both studies reported that the majority of the respondents replied that they should be asked before reusing their information and data should be used only for the consented purpose.

Figure 1 shows the relationship between concepts (Privacy, Trust, Data Access, Research Result and Healthcare Quality) contained in privacy and health outcomes. *Arrow a* shows that privacy and patients-healthcare givers trust are directly proportional. Patients hardly differentiate between healthcare institutions and researchers (mainly working with clinical records). Consequently, they may fail to trust their doctors and even the institution, when privacy is at risk (10). The lack of trust in their healthcare provider could be a cause for patients to self-medicate their illnesses, take traditional medications, tell lies, change doctors, ask their information not to be registered in the EHR, be unwilling to participate in clinical trials and be reluctant to give consent to any public health research (10,11,44). Similarly, minorities refusal to involve in research is also reported because of lack of trust in the researchers (45). Goldman et al. (46) describe the risk of privacy in healthcare as: "*Without trust that their most sensitive health information will be safeguarded, patients are reticent to fully and honestly disclose personal information and may avoid seeking care altogether.*"

Moreover, physicians and healthcare institutions are often unwilling to disclose their patients' data for research with the intention of protecting their patients' or even their own privacy (47). Therefore, lack of trust between patients and healthcare institutions; between healthcare institutions and researchers would affect the research data accessibility and also quality (48) shown by arrow *b* in Figure 1.

Because of the following reasons, epidemiological and healthcare service research require large amount of data across various institutions (1,49,50).

1. Large amount of data is required to achieve strong statistical power
2. Heterogeneity is required to achieve generalizability of research results

3. Disease surveillances require data from large geographical area
4. Rare disease require data from many institutions, as only few cases are treated at individual institutions

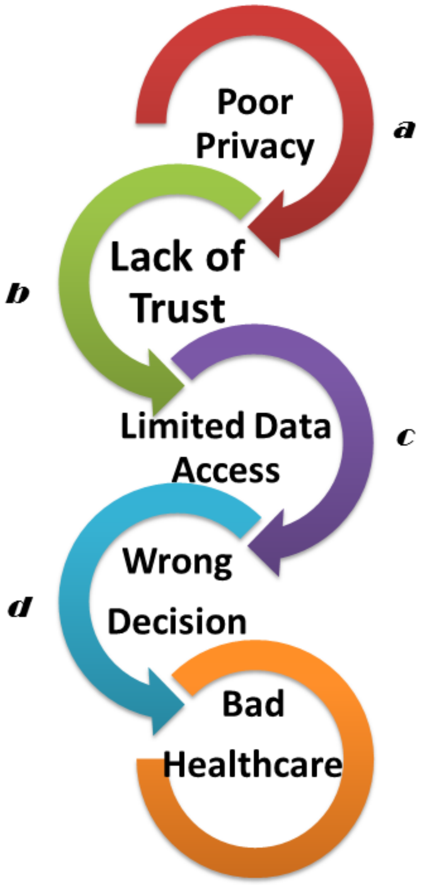


Figure 1. A conceptual model showing consequences of privacy issue in healthcare

Consequently, the amount and heterogeneity of data available for research would result in wrong sampling which could affect the statistical power, which in turn affects the conclusion drawn from the research results (51), as shown by arrow *c* in Figure 1. The consequence of applying inaccurate research results for healthcare decision-making minimizes the quality of care, and could even harm human life, as illustrated by arrow *d* in the Figure 1.

World Medical Association Declaration of Helsinki states that protecting privacy of research participants is one of the basic principles for all medical research (52). Likewise, many other research (1,6,11,13,47) put privacy as the primary concern when it comes to EHR data reuse. Meanwhile, individuals' right to privacy should be balanced with the benefit of the public that can be attained by research (33).

2.2. Privacy preserving techniques

De-identification and cryptographic approaches are the two general privacy-preserving techniques. De-identification is mainly used for centralized approach while cryptographic is for distributed approach (53). Both categories have various implementations for which the high level description is given below.

2.2.1. De-identification

Data de-identification (also called anonymization in some places) is a method of protecting research participants' privacy by removing or modifying personal identifiers. De-identification involves altering (i.e. masking, suppression, randomization, adding noise, and generalization) (29,44) or removing sensitive information in order to reduce the probability of re-identification of individuals (10,29). For example, *HIPAA Safe Harbor* removes 18 identifiers whereas *Limited Data set* removes 16 of the 18 identifiers, except dates and some geographical data (28). However, the advancement in technology and data mining tools make re-identification easier (10,48,55). For example, a research (56) has shown that there was a high risk of re-identification of the US citizens by combining only three fields of information (gender, ZIP code and date of birth). On the other hand, strong de-identification has a potential to decrease data utility (57,58). For example, if all demographic information is removed for privacy reasons (by de-identification), the data can hardly be used for surveillance research. Therefore, de-identification requires a tradeoff between privacy guarantee and data utility.

2.2.2. Secure Multi-party Computation

Secure multi-party computation (SMC) is a cryptography based computing technique for multiple parties to jointly compute on their secret values and at the end of computation only computation results are revealed (59). In SMC, basic analysis (the one that involves sensitive information) is done at the original data source. Consequently, it allows participating institutions to have control on their data.

Compared to de-identification technique (centralized approach), SMC (distributed approach) favors "patient level data stay at their original places". Besides, original data are not affected in the privacy protection process, since SMC do not alter or remove data attributes (53). A very recent article by Dan Bogdanov et al. (60) stated that SMC has a "potential paradigm shift in data protection" because of its capability to perform computation without the need to see individuals' data values. The same article reported results of an interview with SMC end-users from 6 European countries and concluded the following challenges, (1) not seeing the data sets makes the user skeptical about the computation result; and (2) lack of user-friendly tools for performing the data analysis are the major worries even if they are interested in

SMC. To overcome these challenges, the article suggested development of user-friendly environment for end-users and give the possibility to perform descriptive statistics so that they can feel the data on which they are computing.

SMC protocols

There are three generic techniques of developing SMC protocols 1) garbled circuit (61) – a computation technique based on encrypting the computation function; 2) homomorphic encryption (62,63) – a technique that performs calculation on encrypted data without the need to decrypt it; 3) secret sharing (64) – sharing data among multiple parties without the need to know the type of data or type of computation. The performance and complexity of each technique varies based on the specific scenario for which it is implemented (65). For example, Yao's garbled circuit protocol is more efficient for a computation with two input parties than multiple input parties. Or in the case where there is only a single output party, homomorphic encryption can be a better choice though there are other challenges like high computation cost.

Generic SMC protocols are considered not feasible for practical implementation because they are inefficient (66). Goldreich (67) highlighted the need for specialized "application-oriented" SMC protocols for practical use. Specialized SMC protocols are efficient because they are designed to handle specific situations (53). Some of the most widely studied specialized SMC protocols include secure summation (68,69) and scalar product (66). Secure set union, secure set intersection cardinality (70), private permutation (71) and computing entropy (72) are also among the specialized SMC protocols. Furthermore, SMC protocols specifically designed for geometric calculation (73) and data mining (74) also are being developed.

SMC protocols are designed to achieve privacy guarantee against specific adversarial model including corruption strategy, computation complexity and adversarial behavior. The three main categories of adversarial behaviors are (75):

A. Semi-honest (honest but curious) - all the parties follow the rules of the protocol but out of curiosity they might try to learn other parties' private information from the messages exchanged during computation.
B. Malicious – some corrupted parties may arbitrarily deviate from the rules of the protocol to learn private information of other parties.
C. Covert – corrupted parties may arbitrarily deviate from the rules of the protocol but do not wish to be caught cheating.

Often privacy guarantee of SMC protocols is inversely proportional to efficiency and scalability, because of the complex techniques used to ensure stronger privacy guarantee. Therefore, the stronger the security guaranty of a protocol is the less efficient it is. Protocols secure against semi-honest adversaries are known to be more efficient and scalable than others (53). Thus, protocols secure against semi-honest adversaries are sufficient for joint computation among health institutions. Vaidya (76) also has similar argument. As it is one of the technical frameworks of the thesis, secure summation is further studied in *Section2.4.1*.

2.3. Literature Review

2.3.1. Motivation

The motivation for the literature review was to get better understanding of (1) privacy-preserving data reuse research area and identify the knowledge gap that our research question could fill; (2) different privacy-preserving data reuse solutions, and as a result to develop better techniques that solve the research questions.

2.3.2. Method and Scope of the Review

We have searched publications in major journal databases such as PubMed, IEEE Xplore (the online library of the Institute of Electrical and Electronics Engineers), ACM digital library (the online library of the Association for Computing Machinery) and ScienceDirect. Since the concept of EHR data reuse is new, all the reviewed articles are published not earlier than late nineties. In fact, most of them are later than 2008. The literature review was performed in September 2014, so it didn't include publications done afterwards.

The search was done in many steps by using the conjunction words "AND" and "OR" to combine different keywords that mainly characterize the research questions, including *privacy, security, confidentiality, health data, clinical data, health records, EHR, medical records, EMR, clinical records, distributed health data, health research, statistical analysis, statistical processing, query, sharing, data reuse, secondary use, research infrastructure* and *research network.*

In total, 1856 papers were identified from the four databases. In this search, papers written in other languages than English and studies done on animals (Veterinary Medicine) were excluded. 62 duplicate papers were removed from the total result, where 1794 papers left for screening. As illustrated in *Figure 2*, the screening was performed in multiple rounds by using different inclusion and exclusion criteria as described below.

1. Screening by "Title"

In this first round, we went through the titles of the 1794 papers and selected 211 papers those focusing on privacy preserving health data reuse in general. The excluded 1583 papers were those papers with no direct emphasis on privacy. For example, focus on data quality, interoperability and other issues in health data reuse.

2. Screening by "Abstract"

In this round, we looked into abstracts of the selected 211 papers and further assess papers that are focusing on the techniques for preserving privacy in health data reuse. Hereby, we included 90 papers and excluded 121 papers, which are focusing on theoretical aspects of privacy, such as papers reporting study results by using privacy preserving health data reuse infrastructures, review papers.

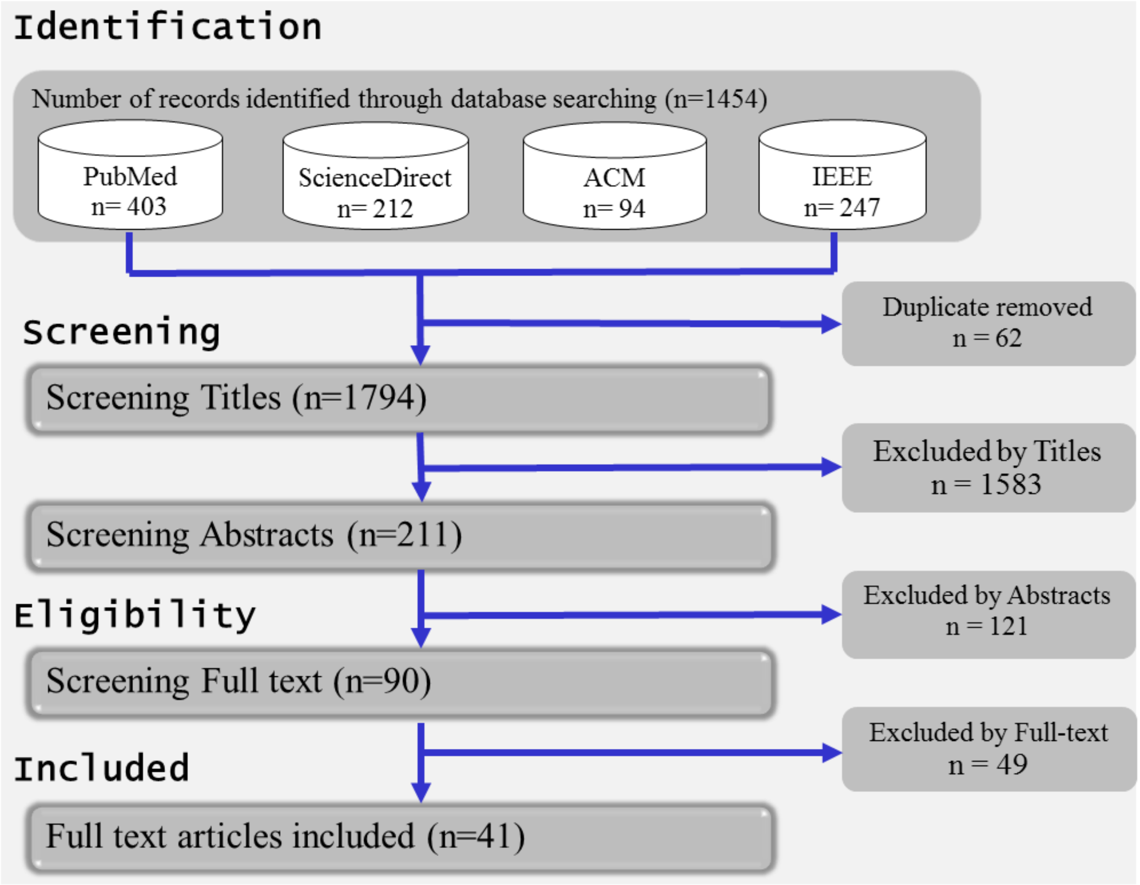

Identification
Number of records identified through database searching (n=1454)
PubMed n= 403
ScienceDirect n= 212
ACM n= 94
IEEE n= 247
Screening
Duplicate removed n = 62
Screening Titles (n=1794)
Excluded by Titles n = 1583
Screening Abstracts (n=211)
Eligibility
Excluded by Abstracts n = 121
Screening Full text (n=90)
Included
Excluded by Full-text n = 49
Full text articles included (n=41)

Figure 2. Summary of the review results in "PRISMA statement" format (77).

3. Screening by "Full text"

The final round was done by reading full text of the 90 eligible papers (see Appendix A). In this round, 41 papers were selected (included) as the most relevant paper for the thesis. The inclusion criteria were papers with primary focus on privacy preserving techniques in a distributed approach (i.e. individual level data cannot be moved outside the health institutions). The 49 excluded papers were those papers emphasizing privacy preservation in a centralized approach (i.e. collecting data by de-identification and consenting mechanisms). Some of the selected (included) papers are presented in the "related work" section and others are referenced in different parts of this thesis. The excluded papers are also used as background knowledge for the thesis.

Figure 3 indicates the increasing focus of publications on privacy aspects of health data reuse. This might be due to the increasing interest to unlock the power of enormous data being available or increase in privacy concerns.

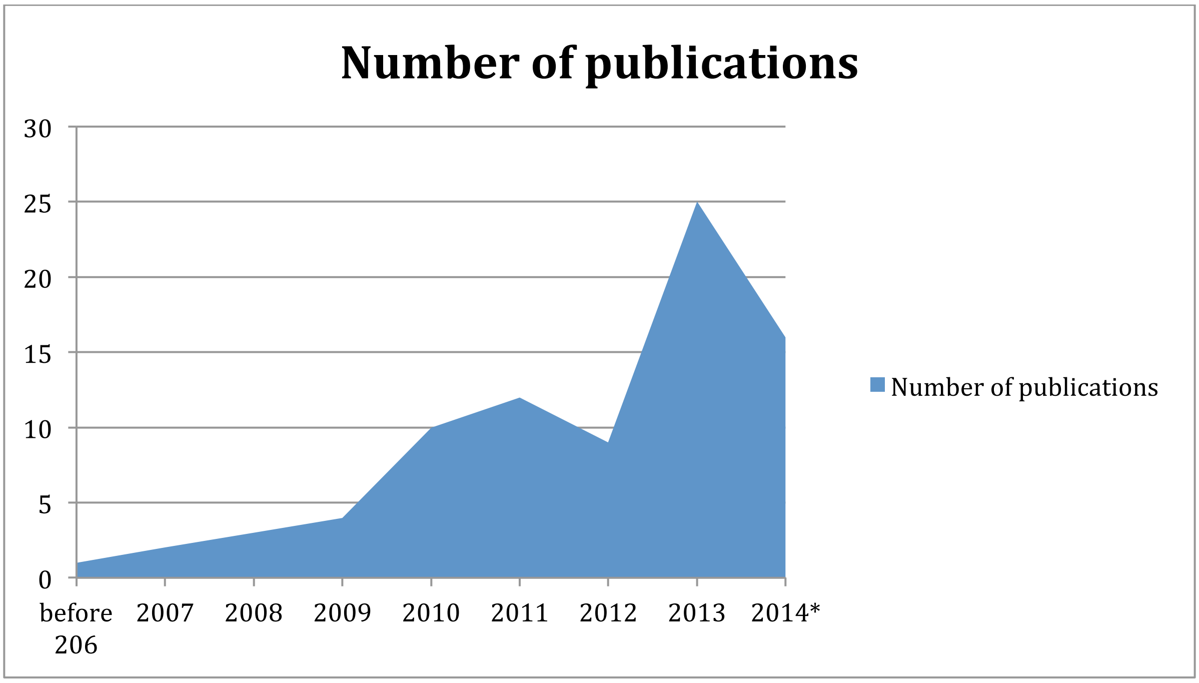


Figure 3. Trends of publications focusing on privacy concerns of health data reuse by year.

**Note that the statistics shown in the figure didn't include papers published after September 2014.*

In general, the literature review process was a learning opportunity that offered us a new insight into the research area and introduced us with interesting and innovative works that are relevant to this thesis. The results of the literature review are described in the following subsection.

2.3.3. Related Work

The idea of privacy-preserving health data reuse for research purpose is in its early stage (78). However, lots of efforts are being made on creating suitable infrastructures for reusing medical records for research (14,16–18,79,80). Some projects use a centralized approach, a large central data warehouse where data from different health institution are collected and made available for researchers. Alternatively, other projects use the distributed approach, which enables statistical analysis without moving data outside the health institutions.

Sittig et al. (81) reported a survey of distributed research networks that are developed to facilitate reuse of distributed health record. The tools described in the article include SPAN, WICER, CER-HUB, RPDR, INPC COMET-AD, and SCOAP-CERTN. Except SPAN and CER-HUB, all of the tools use the centralized approach.

The major query tools and distributed research networks that use the distributed approach are discussed below. These tools have wide functionalities than this thesis. However, the discussion is limited to techniques used for different steps of privacy-preserving statistical computation, such as common data model, and implemented statistical functions.

1. SHRINE

Shared Research Information Network (SHRINE) (82) is a tool built to query clinical data repositories based on the data model and functionalities of i2b2.

I2b2 - Informatics for integrating biology and bedside (i2b2) (79) is a software with a set of tools that helps researchers explore medical record to undertake research while preserving patient privacy. It also provides the querying capability using a web based user interface.

The i2b2 software has two main use cases. The first one is creating patients' data sets based on some research criteria while the second one is further exploration of the selected data sets. Therefore, authorized researcher can use the query tool to create specific set of patients or to get the total count of demographic information of the selected data set. For the sake of preserving privacy, query results are de-identified in some way and a small random numbers are added to the actual result before displayed to the users. The querying capability of i2b2 is proven to be successful in an enterprise level database with millions of records.

i2b2 has its own data model and uses the "star schema" design for clinical research chart. The star schema has four tables (patient_dimension, concept_dimension, visit_dimension, observer_dimension) and they are related to a central table called observation_fact.

As depicted in Figure 4, SHRINE broadcasts query to different data repositories and uses central component (SHRINE Query Entry point) to combine the results from each participating data repositories before displaying it to the user on the *SHRINE Web client*.

In order to achieve semantic interoperability among the multiple heterogeneous sites, SHRINE performs the following four consecutive steps:

I. Extract local clinical data – extract data from clinical database into local research database,
II. Map local concept codes to standard concept codes – key-value pair mapping of local concepts with the four standard categories of clinical concepts,
III. Group concepts using medical hierarchies – grouping and maintaining relationship of concepts using standard medical hierarchies and
IV. Adapt query to use local concepts codes – making incoming queries to use local concepts

Many projects are taking advantage of the i2b2 data model and the query capability of SHRINE to build their own research infrastructure. I2b2-SSR and **SCILHS** are two examples.

I2b2-SSR (83)**-** It is a project in US designed to build self-scaling registry technology for collaborative data sharing, based on the widely adopted i2b2 data warehousing framework and the SHRINE peer-to-peer networking software.

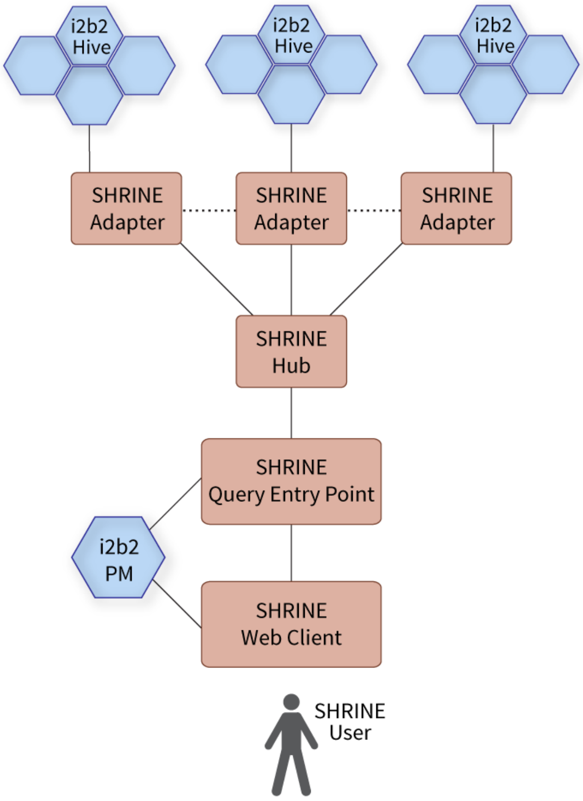


Figure 4. The high-level architecture of SHRINE(84).

Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS) (85) –The goal of this project is to create research agendas, approval studies, identify participants from diverse populations, notify them about research, enroll them in trials, study the cohort with ongoing bidirectional communication, and return research results. To achieve these goals, SCILHS uses the i2b2 common data model and SHRINE to query multiple sites at a time.

2. PopMedNet

PopMedNet (17) is a software designed to facilitate analysis of distributed health data for different types of research. This tool has features such as privacy preserving techniques, menu driven data analysis process, scalability, and support different data models.

The architecture of PopMedNet comprises of two main components single portal and many DataMart Clients. The portal is the entry point where the user creates a query and gets the results; and it also support security and other network administrations. Whereas, the DataMart Clients are the components located at each data sources and responsible for processing queries and return results back to the portal in a secure way.

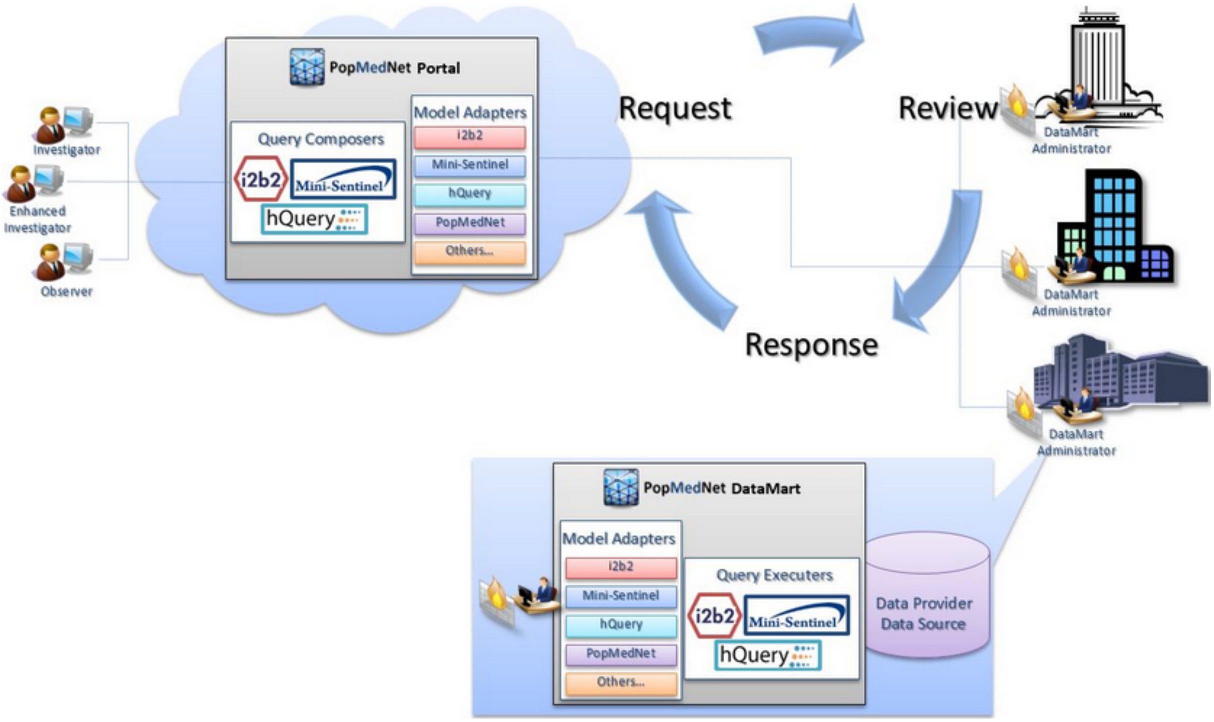


Figure 5. The high-level architecture of PopMedNet (86).

As shown in Figure 5, the current version of PopMedNet supports variety of data models such as i2b2, HMO Research Network's (HMORN) Virtual Data Warehouse, the Mini-Sentinel Common Data Model, and the ESP data model. Besides, it gives the flexibility to implement any other data model (87).

Several distributed health research infrastructures use PopMedNet to query distributed health databases. Some of the major research networks are:

Query Health (88) – an initiative for building a nationwide architecture for distributed, population-level health queries across diverse clinical systems in the US. This project makes use of PopMedNet, i2b2 and hQuery. While PopMedNet is used for distribution of queries and getting results in a secure manner, i2b2 and hQuery are used for processing data analytics.

Mini-Sentinel (89)– a project sponsored by the US Food and Drug Administration for making active surveillance system by analyzing combined EHR data from different health institutions. This pilot project uses PopMedNet for secure distribution of query and cohort selection from the distributed EHR.

Other projects such as HMORN (90), PCORNet, comparative effectiveness research (National Patient-Centered Clinical Research Network), the National Institutes of Health's Health Care Systems Research Collaboratory Distributed Research Network (biomedical research), and ESPNet (public health surveillance) also use PopMedNet together with other research infrastructures to create a suitable and secure research environment (91).

3. SCANNER

SCAlable National Network for Effectiveness Research (SCANNER) is a network to facilitate a secure, and scalable research among distributed health institutions in the US (18). As Figure 6 shows, the major components of the network include: portal, registry, Master node, and Worker nodes. The work starts from the portal where the researcher authenticates and composes a query using the provided selection controls (stored in the registry) that tell the available data sets, computational algorithms and remote nodes. The portal then sends the query to the Master node which in turn issues the query to the worker nodes found at remote sites. After execution of the query, the worker nodes return results to the Master node for final computation. The Master node then sends the result to the portal where the researcher can see it.

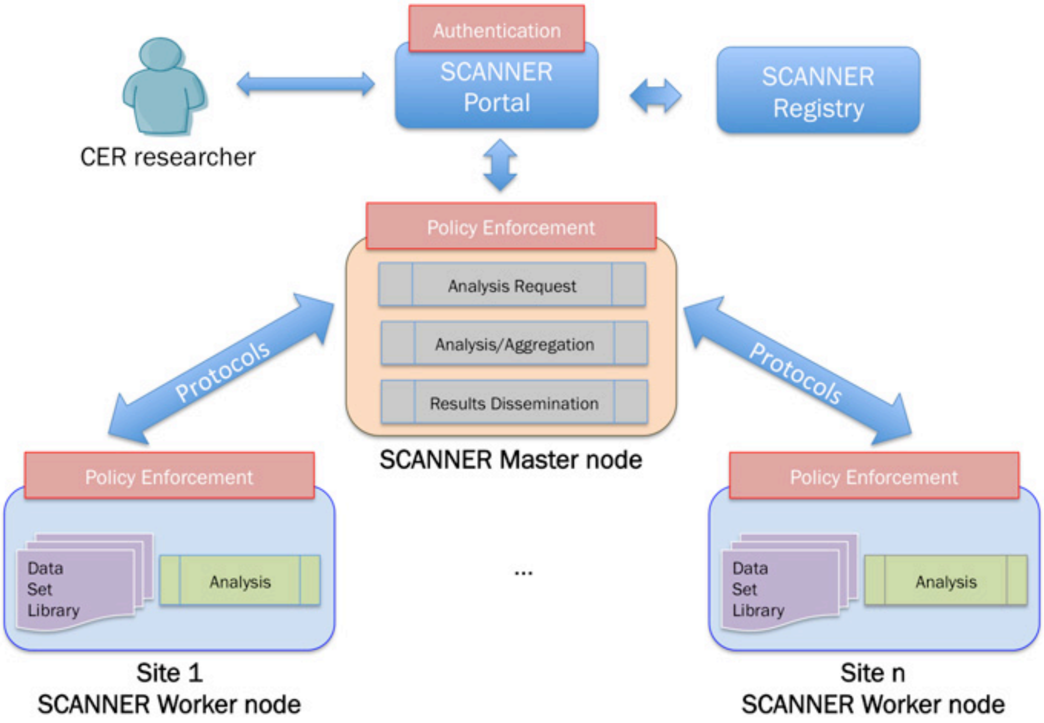


Figure 6. The conceptual architecture of SCANNER (92)

To maintain both syntactic and semantic interoperability, SCANNER uses Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) version 4.0 (93).

4. ***EHR4CR***

EHR4CR (Electronic Health record for Clinical Research) is an European project aimed at creating a suitable platform for reuse of EHR data across healthcare institutions in multiple countries for research (34). To maintain semantic interoperability between the heterogeneous health information systems, EHR4CR proposed shared conceptual reference model (EHR4CR information model), which is HL7-based UML model annotated with the concept of shared EHR4CR terminology (9). As shown in Figure 7, each health institution run data queries against its database and store query results in a separate database and then it is available for further data processing by the user. EHR4CR platform uses Extract Transform and Load (ETL) process to support different databases with different data formats. It also ensures end-to-end confidentiality by implementing WS-Security encryption for SOAP messages and XML Encryption for REST services exchanging XML messages (94).

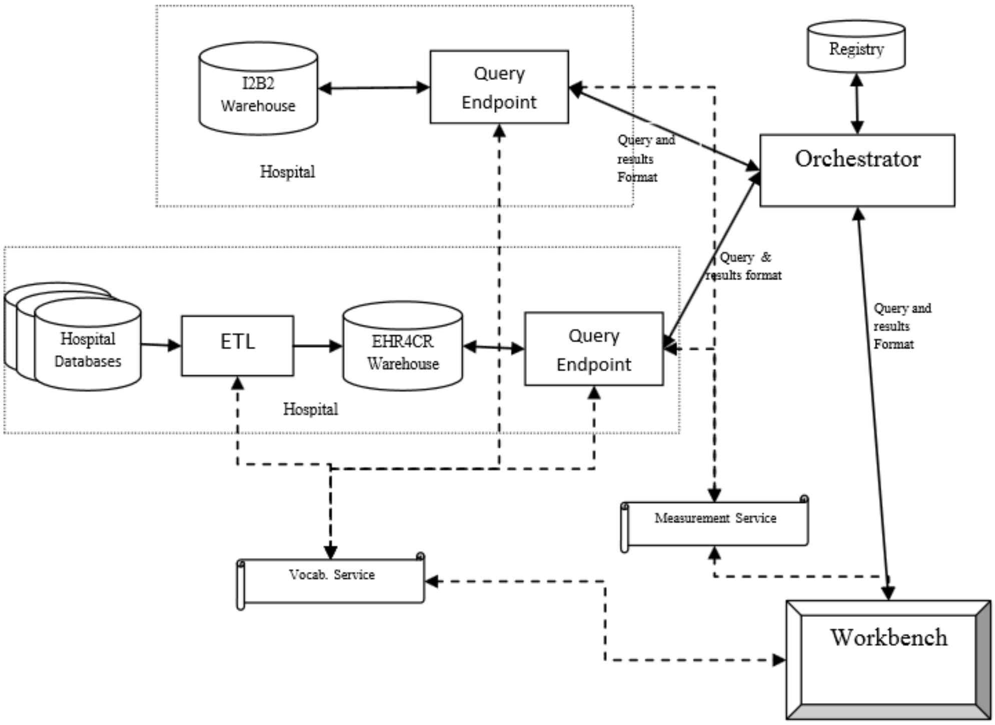


Figure 7. EHR4CR: PFS Services and components interaction(94)

5. ***SAFTINet***

SAFTINet stands for **S**calable **A**rchitecture for **F**ederated **T**ranslational **I**nquiries **N**etwork (SAFTINET) (95). As shown in Figure 7, SAFTINet components are: (1) a query system composed of web-based Query Portal (QP) and Federated Query Processor (FQP) for data request from distributed nodes, (2) Translational Informatics and Data management Grid to facilitate communication between the query system and partner the nodes, (3) OMOP-SAFTINet interface transformation adapter (ROSITA) for data transformation to a common data model (OMOP CDM V4) and (4) Partner Node with data formatted as a HIPAA-compliant limited data set in the OMOP CDM V4 format.

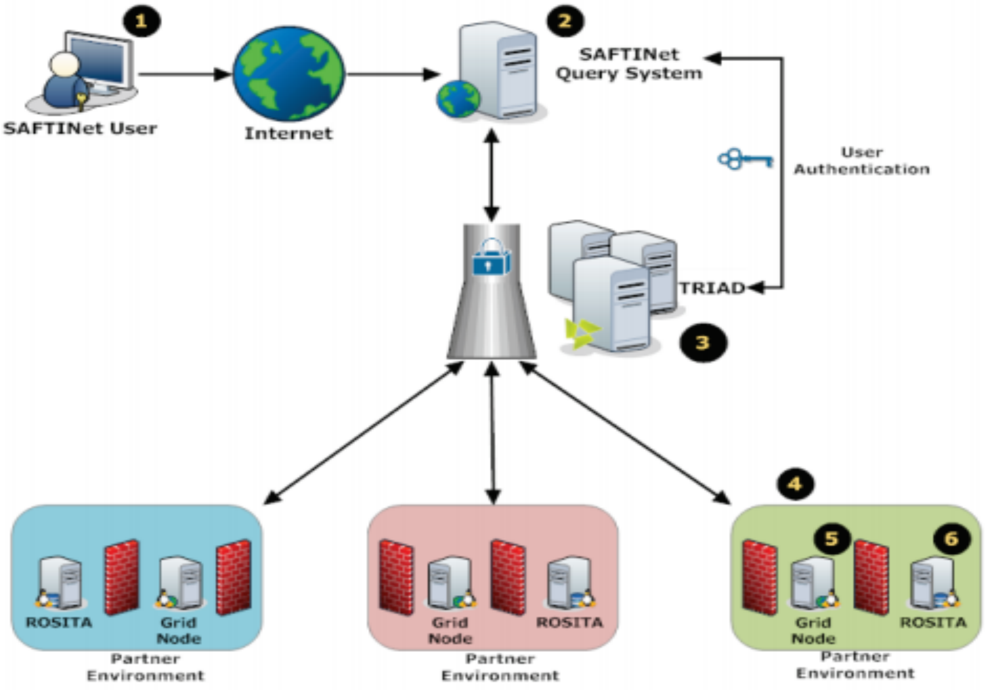


Figure 8. SAFTINet Infrastructural overview (95)

The FQP is located at the central and is responsible for forwarding queries to each grid partners behind firewalls; and aggregating results (from each grid partners) to be displayed to the user through the QP.

Table 1. Summary of the major query tools and distributed research networks.

#	Name	Data Model	Whose privacy is protected?	Query supported
1	SHRINE	i2b2 data model	Individuals'	Count
2	PopMedNet	HMORN VDW, Mini-Sentinel Common Data Model, ESP data model	Individuals'	Count, Prevalence, Incidence, Cohort selection
3	SCANNER	OMOP data model	Individuals' and Institutions'	Cohort selection, Descriptive statistics (Count, MIN, MAX, AVERAGE, VAR, STD, SUM MEDIAN), Logistic regression
4	EHR4CR	EHR4CR information model	Individuals'	Count
5	SAFTINET	OMOP data model	Individuals'	Count

Table 1 summarizes major statistical query tools and distributed research networks such as SHRINE (82), PopMedNet (17), SCANNER (18), EHR4CR (34) and SAFTINET (95) which implemented statistical analyses on distributed health data. As shown in the table, all tools except SCANNER only support statistical count. Besides, they disclose individual institution level count, which does not address institutions' privacy concerns. In contrast, SCANNER supports more statistical analyses and has implemented computation techniques that release aggregated statistics of multiple institutions' data, to protect individual institutions privacy.

2.4. Technical Framework

In this subsection we describe the technical framework of the thesis, including the review of secure summation protocol, the Snow system, and the openEHR platform which are basic components of the prototype developed in the thesis.

2.4.1. Secure Summation

Secure summation is one of the most commonly studied SMC protocol and used as a building block for several secure computations (96). The basic principle of the protocol is to perform secure addition of distributed data without revealing individuals' values. As discussed in (60,97), many statistical problems can be decomposed into a set of sub-computations of summation form. Decomposition of statistical functions' implemented in this thesis is described in detail in chapter 5.

Secure summation protocols are designed using different techniques, such as secrete sharing (98,99**)**, Homomorphic encryption (100) and adding random number on a secret value. The most common ones are revised in the following paragraphs.

Simple Secure sum - secure summation can be implemented based on adding random number to the secret value before sending to another institution during the summation and subtracting the random values from the final value. The protocol given below explains this concept.

Karr et al. (101) showed a simple secure sum protocol shown in *Figure 9.* A coordinator node C 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 $R + S_1 + S_2 + \dots + S_n$ to find the true sum of the secret values $S_1 + S_2 + \dots + S_n$.

The protocol is efficient because it: (1) uses a simple technique; (2) involves only *n* number of communication for *n* 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 node $i + 2$ collude to find a private value of node $i + 1$. In semi-honest adversarial model, collusion of the two neighbors of a node is a common problem of this kind of secure summation protocol (102).

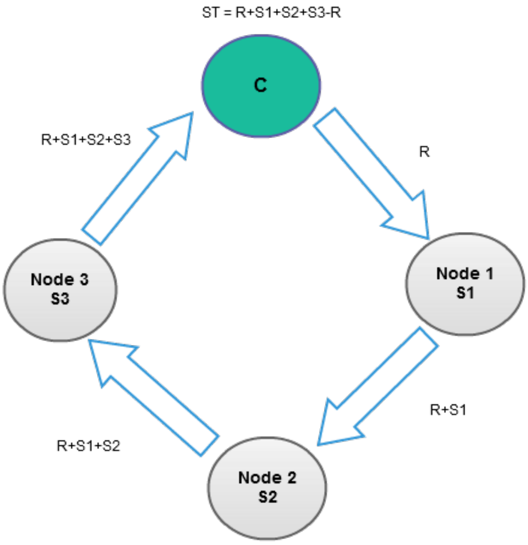


Figure 9. Simple secure sum protocol

SINE (Secured Intermediate iNformation Exchange) - Shuwang et al. (103) implemented random number based protocol with improved collusion resistance. The computation starts with a coordinating server sending a random number R_c to the first node. The first node then adds its own random number R_1 on R_c and passes it $(R_c + R_1)$ to the second node which does the same and sends it $(R_c + R_1 + R_2)$ to the next node. Finally, the last node adds its own random number and forwards it $(R_c + R_1 + R_2 + \cdots + R_n)$ to the coordinating server. Then, the coordinating server 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, every node sends the sum of its private value and its own random number $R_1 + S_1$, $R_2 + S_2, \dots,$ and $R_n + S_n$ directly to the coordinating sever. The server then sums up these values and subtracts the sum of random values $((R_1 + R_2 + \cdots + R_n) + (S_1 + S_2 + S_n)) - (R_1 + R_2 + \cdots + R_n)$. Therefore, this protocol minimizes the collusion problem, while the number of required communication increased to 2n-1 for n number of nodes.

Secure sum with Secret Sharing - another way of computing secure summation is using secret sharing scheme (99) as a building block. In this scheme, the protocol has two phases. In the first phase every participating node divides its private value into n secret share values, keeps one to itself and distributes the rest to the other nodes. Only significant subsets of nodes can reconstruct the secret value. In the second phase, every node adds the shares received from the other nodes and the share it kept for itself, and sends to the coordinating node. The coordinating node subsequently computes the total sum by adding the values received from the nodes, without knowing private values of the nodes. Examples of this implementation is presented in (98,102). The main idea of dividing a value into random shares is to avoid having the same neighbors twice for each cycle. As a result, collusion between neighbors of a node decreases. Especially (102), showed the best collusion resistant secure sum protocol such that privacy of a node can be violated only if n-1 number of nodes collude.

Secure sum with homomorphic encryption – Homomorphic encryption is a form of encryption that uses homomorphic properties of cryptosystems to enable computation out on ciphertext without decrypting. Secure summation protocols are implemented using homomorphic encryption as a building block. Thus, such protocols are known to have a strong security guarantee. Drosatos et al. (100) showed how different statistical equations can be implemented by secure summation protocol with homomorphic encryption. However, such protocols have higher communication cost, since ciphertext is much larger than plaintext, and high computation complexity as a result of the encryption and decryption. Therefore, the protocols are less efficient especially when the number of nodes increases. Besides, using partial homomorphic encryptions are limited to compute one computation such as multiplication or addition over the encrypted values without decrypting (65). And the full homomorphic encryption techniques are extremely computationally expensive and not ready for practical applications (104).

Secure sum with penalty – the game based secure sum approach introduced in (105) has the idea of penalizing nodes which get caught trying to collude. The penalty could be exclusion from the computation or increasing the computation and communication cost. However, there is also a high probability for the cheating (colluding) nodes not to get caught since there is no defined way of tracing them. Such protocols are secure against a covert adversary who doesn't want to be caught cheating, for main reasons including status, including legal and societal status.

Table 2. Summary of the performance and privacy of different secure sum protocols.

No.	Secure Sum protocol	Communication overhead (bit)	Computation complexity	Reference
1	Simple secure sum	$b_p n$	O(n)	(101)
2	SINE	$b_p(2n+1)$	O(n)	(103)
3	Secret sharing	$b_p n(n-1)/2$	$O(n^2)$	(98,102)
4	Homomorphic Encryption	$b_p n+ b_e(n-1)$	O(E*n)	(100)

Table 2 depicts the comparison between the secure summation protocols discussed above with regard to communication overhead and computation complexity. In this table, "*n*" represents the number of nodes participating in the computation while "b_p" = bit length of private value and "b_e" = bit length of encryption information.

In summary, privacy and performance (efficiency) are inversely proportional in this context. Consequently, it is difficult to pick a perfect protocol that can handle every computation in a proficient manner. We might require a little compromise between privacy and computation efficiency in order to get acceptable amount of work done. Meanwhile, mechanisms to improve the efficiency of SMC protocols are being developed. For example, Yigzaw et al. (106) suggests by using the concept of parallel computation (i.e. divide a task into concurrent sub-computations among peers and aggregate the final result), scalability and efficiency of SMC can be enhanced.

2.4.2. Snow System

The Snow system (107) is a mobile agent based peer-to-peer system that enables trans-institutional access to reuse of patient data. Snow mainly works on *disease surveillance* by extracting anonymized data from EHRs of GP offices, hospitals and microbiology laboratories. The extracted data is used to inform GPs and patients about possible outbreak and disease forecast in a geographically and timely manner.

The Snow system is a XMPP based relay network of servers between the snow system installation in health institutions (EHR systems) and Healthnet[4]. This relay network enables global message delivery between health institutions. As it is a part of the Snow project, the system developed in this thesis uses infrastructure and technologies used in Snow system such as authentication, message broadcasting, message encryption and servers.

2.4.3. OpenEHR

OpenEHR is an open standard specification to attain semantic interoperability among electronic patient records (EHRs). OpenEHR uses clinical knowledge resources such as archetypes and templates to represent the health data. Archetypes are general concepts containing a maximum data set that is pruned and adapted to our local use by means of templates. A template is the mechanism to combine several archetypes for the local use.

OpenEHR uses two-level modeling approach, the information level and knowledge level. In this approach, the information level contains a stable reference information model, which is not likely to change over time while the knowledge level defines constraints over the information model that allows specifying clinical concepts in the form of archetypes and

[4] https://www.nhn.no/english/Pages/about.aspx
A national network that connects Norwegian healthcare institutions together for the purpose of information exchange

templates. Since the second level involves openEHR archetype, systems deployed based on openEHR archetype can exchange health information regardless of the system environment and programming language they use (26)(108). Therefore, openEHR is a future proof platform for this thesis.

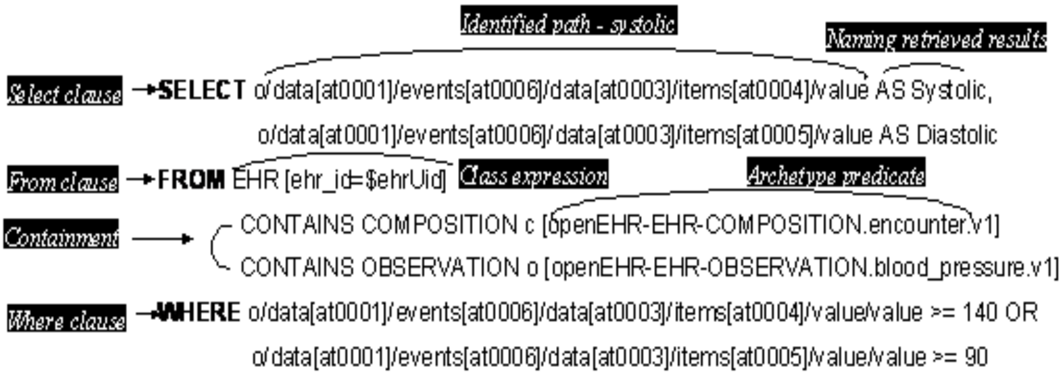
2.4.4. OpenEHR CKM

The openEHR Clinical Knowledge Manager (CKM) is an international, web-based clinical knowledge resource created by a group of professionals including clinicians and computer scientists for the purpose of sharing application and message independent health information (109).

Likewise, Norwegian CKM (110) is a Clinical Knowledge Manager hosted by Norwegian national eHealth program. Norwegian CKM is being developed to publish more archetypes for national use. Some of the archetypes are approved for usage and are listed in the archetype registry.

2.4.5. AQL – Archetype Query Language

AQL is a query language developed to perform queries on openEHR-based EHRs. It was first developed by Ocean Informatics and was initially known as the EHR Query language (EQL) (111). AQL is neutral to any EHRs, programming languages, and system environments and only depends on openEHR Reference Model (RM) and the openEHR clinical archetypes. The specialty of AQL is that it expresses the queries at the archetype level, i.e. EHR information standard level rather than at the persistence schema or application level, which is not the case for other query languages such as SQL or HQL. This makes AQL queries reusable across different boundaries and systems (112). AQL makes the queries independent of the underlying technology. Thus, we can move the queries from one system to another without caring about the technologies the platforms are implemented; e.g. SQL, XML.



```
SELECT o/data[at0001]/events[at0006]/data[at0003]/items[at0004]/value AS Systolic,
       o/data[at0001]/events[at0006]/data[at0003]/items[at0005]/value AS Diastolic
FROM EHR [ehr_id=$ehrUid]
    CONTAINS COMPOSITION c [openEHR-EHR-COMPOSITION.encounter.v1]
    CONTAINS OBSERVATION o [openEHR-EHR-OBSERVATION.blood_pressure.v1]
WHERE o/data[at0001]/events[at0006]/data[at0003]/items[at0004]/value/value >= 140 OR
      o/data[at0001]/events[at0006]/data[at0003]/items[at0005]/value/value >= 90
```

Figure 10. . Example of AQL query to retrieve Systolic value greater than 140 and Diastolic value greater than 90 (112).

AQL uses the basic commands such as SELECT, FROM, WHERE, ORDER BY, and TIME WINDOW (to support queries with logical time-based data rollback) (112). The following example (Figure 10) shows a typical AQL query to retrieve systolic and diastolic blood pressure value from any archetype-based EHR that has the Blood Pressure archetype.

2.4.6. OpenEHR Repository

As a representative of the openEHR repository we used Think!EHR. Think!EHR (113) is a platform developed by a company called Marand[5] based on state of the art technology and industry standards such as openEHR, IHE and HL7. Think!EHR offers a range of functionalities including storage, management, query and exchange of structured EHR data. The Think!EHR platform conforms to the latest version of openEHR specifications and is built on java technologies. Moreover, it provides visual tools for form generation and query building, API used to build applications and developer's toolkit with code samples included. Think!EHR uses AQL as query language and provides easy to use, web based EHR querying tool called "EHR Explorer", which displays the query result in a tabular form.

2.5. Summary

This chapter described the overview of the-state-of-the-art in the field of EHR data reuse and statistical analyses on distributed health data. We discussed the basic knowledge about privacy and the necessity of privacy in EHR data reuse. Following that, privacy preserving techniques (i.e. consent, data de-identification and secure multi-party computation (SMC)) used in healthcare research are discussed. Then, we elaborated the discussion on SMC and compared different secure summation protocols. The comparison was mainly on the basis of privacy guaranty, communication overhead and computation complexity.

We also presented the methods used for literature review on privacy-preserving health data reuse. The literature review covered a range of articles published in different journals. Then, the works that are closely related to this thesis have been briefly presented.

The technical frameworks used in the prototype development, such as secure summation protocol, openEHR, Archetype Query Language, and related tools have also been discussed.

[5] http://www.marand-think.com/

Chapter 3 Materials and Methods

3.1. Research Paradigms

Agile software development methodology are a set of methods and tools used for software development (114,115). During the last decade these methods have become popular. The main advantages of agile methodologies over traditional methodologies are they provide capability to adapt to changes in requirements and to learn from development experiences among a team. As a result, they are believed to increase flexibility and productivity of software development, and quality of the software (116).

In this thesis, we have used agile methodology to develop a prototype system (we named it *Emnet*) that implements the solutions developed to the specified problems.

There are different agile methodologies. The most widely used methodologies are XP, and Scrum. Often agile methodologies give effective practice for the different phases of the development process. Therefore, they are considered complementary to each other (116).

Scrum is a simple and adaptable framework (116). In addition, in the Snow project, where this thesis is part of, we have experience in using Scrum. Therefore, in this thesis we have chosen Scrum among other agile methodologies.

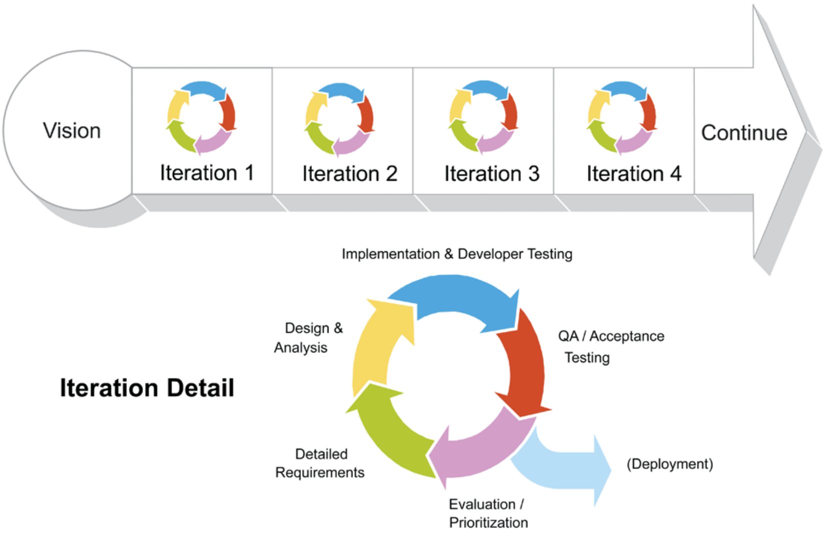


Figure 11. Scrum development activities into iterations (117).

As shown in Figure 11, the basic principle of scrum is dividing the whole task in to small iterations called sprints. Every sprint has a sprint goal. Therefore, at the end of each sprit, a part of the whole product is developed and tested. Each sprint builds on previous sprints. Developing one piece at a time boosts creativity and enables to incorporate new requirements and get feedback that helps to develop the right product. One of the agile manifestos is sprint duration “from a couple of weeks to a couple of months, with a preference to the shorter time scale” (115). In the thesis, therefore, sprints are time boxed to two weeks.

3.2. Data Preparation Methods

Basically, the input of *Emnet* is openEHR data. However, we don’t have wide implementation of openEHR in Norway. Consequently, we created data to test the developed prototype based on the scenarios given by the potential users of *Emnet*. The detail process of the test data preparation is explained in chapter 5.

3.3. Materials

The following tools and programing languages are used for the prototype developed in the thesis:

OpenEHR CKM: For openEHR data generated to test the prototype, we have used openEHR CKM to navigate and download archetypes that represents the clinical concepts required in the data.

OpenEHR Archetype Editor: When the archetypes in the openEHR CKM were not sufficient for our use, the editor was used to modify or create new archetypes.

OpenEHR Template Designer: It is used to design templates from a set of archetypes.

OpenEHR Information Repository: We used the Think!EHR platform to store the openEHR compatible data set used in the development and testing of the prototype. Think!EHR is a platform that allows to persist EHR extracts compliant with openEHR and query them using the Archetype Query Language (AQL).

Query languages

Emnet involves various rounds of query execution on two different data formats. The reason for our choice of two data formats is explained in the following chapters. Therefore, we have used two query languages, one for each data format.

> ***AQL:*** It is used to query openEHR compatible data stored in Think!EHR.
> ***SQL:*** It is used to query relational databases.

HTML5: It is a standard programing language used to describe and present the contents of web interfaces.

Cascading style Sheet3 (CSS3): It is a styling language used to describe the look and feel of HTML codes.

JavaScript: It is used to implement client-side scripts to make interactive and dynamic web interfaces.

Java Development Kit (JDK): Java development kit is used as a basic development environment tool in this thesis. We preferred java because the Thinker!EHR platform is built on Java environment. Besides, as Java is a platform independent language, it gives us the flexibility to run the prototype on different operating systems (118). In addition, the prototype is part of the Snow system, which is implemented in Java and we have more Java experience that other languages.

Eclipse: It is an Integrated Development Environment (IDE) that facilitates the coding process of java programs. Eclipse was chosen because it is easy to use and we have experiences.

XMPP: The *Extensible Messaging and Presence Protocol* (XMPP) is an open technology for real-time communication. XMPP uses the *Extensible Markup Language* (XML) for exchanging information. The reason for using XMPP is explained in chapter 6.

Openfire Server: Openfire is an instant messaging server that uses XMPP server written in java. Openfire server was chosen because it is a very popular XMPP server and it is also used by the Snow system (107).

Strophe: It is an XMPP extension library used to bind the communication between a web-based clients (which uses http protocol) and an XMPP server (which uses XMPP protocol) (119).

Smack (XMPP extension): It is an XMPP extension Java library used to make XMPP client in the prototype components implemented in Java (119)

3.4. Critiques of the methods used

Appropriate user involvement is crucial for the success of a project. As more users' involved in the development process, the system usability increases. In this thesis continuous user involvement was difficult. Therefore, we only used Personas (see section 4.2) and the literature review result to get insight on the system requirements.

Furthermore, *Emnet* needs testing in real life environment and performance evaluation. The performance evaluation of the developed techniques would make it more acceptable and reliable.

3.5. Summary

This chapter described the research tools and paradigms used in this thesis. The first section described the agile methodology used to develop the prototype solution to the specified research problem. The next section presented tools and programing languages used in the development process. Finally, critiques of the methods used in this thesis were discussed.

Chapter 4 Requirements Specification

The objective of this chapter is to identify, analyze, and document requirements of a distributed privacy-preserving statistical computing system on EHRs data. The chapter presents the overall description, assumptions, dependencies, and potential users of *Emnet*. Then, the source of the requirements; rationale for the choice of Persona as an alternative for user involvement in the design and development process; and requirements specification are described.

4.1. System Description

Let us assume a set of health institutions' have EHRs data and want to share their data, while protecting the privacy of their patients and their institutions. We are building a system that enables privacy-preserving statistical computing on the distributed data. *Emnet* enables the following two main tasks:

1. Data preparation

In traditional statistical software, such as R[6], SPSS[7] and Stata[8] the users collect data set of their requirements from different sources and mainly use the software to run statistical computations. In contrast, *Emnet* should enable users to specify inclusion and exclusion criteria for the data set of their requirements from enormous amount of data distributed across institutions.

2. Statistical computation

Unlike the traditional approach where the statistical software and data are on the same computer, in the context of this thesis the data are distributed and collecting outside the institutions will be privacy violation. Therefore, *Emnet* should enable users to perform statistical computations on the data set created using data preparation functionality, while protecting the privacy of patients and the health institutions that owns the data set.

In practical use of *Emnet*, there should be a way for the health institutions to control whom to share data, which data to share, and what analyses will be run on the data. In addition, ethical board approval might also be required. However, these aspects are outside the scope of this thesis.

Users

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

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

[8] http://www.stata.com/

Similar to traditional statistical software, *Emnet* can be useful for a wide range of healthcare related professionals. However, the main users include:

a. Epidemiologists
b. Clinicians (e.g. general practitioners (GPs))
c. Clinical researchers

These users can query patients' data and perform statistical computations on data distributed across institutions. However, clinicians can also perform privacy-preserving computations on data at their local institution, assuming that they are authorized to access local data at their institution.

In this thesis, the term "user" is used to refer to any of these users of *Emnet*.

Constraints

In addition to assumptions described in section 1.3, the following assumptions are made:

- Since *Emnet* is part of the Snow system, it will reuse functionalities already implemented in the Snow system. Therefore, functionalities such as user authentication, encrypted communication channel, and monitoring and management of the components running across health institutions are not included in this thesis.
- "Think!EHR" platform cooperates with *Emnet* by providing interface to access the openEHR repositories and facilitating query execution during research data preparation.

4.2. Sources of requirements

We had unstructured meetings with two representative users of *Emnet* to discuss the requirements and use case scenarios of the proposed system. These representative users are both medical doctors who practiced/practicing as a general practitioner. In addition, one of them is experienced epidemiologist and the other one primary care researcher. They work at Norwegian Centre for Integrated Care and Telemedicine (NST).

Appropriate user involvement throughout the design and development process is crucial for the success of a project. In this thesis continuous user involvement was difficult. Therefore, we have used Personas (120,121). It is a powerful complement to other user-centered methods. Yet, used alone it can aid development team to understand and identify the target users and also aid in design and development. In addition, it also provide shared basis for communicating users need (120,121). We developed three personas based on the meetings with the two representative users and have been refined throughout the Scrum sprints.

The other sources of requirements are also the ideas collected from the related works described in the literature review section of chapter 2. Those works helped us to understand the functional and non-functional requirements of such systems.

Because of the iterative nature of the agile methodology used in this thesis, the requirements specification process has been under continuous modification throughout the design and implementation of *Emnet*.

User Personas and Characteristics

Personas of epidemiologist, GP, and clinical researcher are developed using the Microsoft Solution Framework (MSF) agile Persona Template (120). The personas also had given names in order to refer them easily later in the thesis.

Table 3. Persona of Epidemiologist.

Name: Dr. Maria	**Status and Trust Level:** High
Role: Epidemiologist (health service researcher)	**Gender:** Female

Knowledge, Skills and Abilities: Dr. Maria knows how to use scientific rules and methods, logic and reasoning to solve problems. She also has the basic knowledge of using statistical and mathematical software to perform research analysis. She also knows how to collect scientific evidences for interventions, harm reduction, risk management, or prevention strategies.

Goals, Motives and Concerns: to perform statistical computations on EHRs data in order to understand how often diseases occur in different groups of people and measure health outcomes.

Usage Patterns: Dr. Maria is a regular user of *Emnet*. She uses it throughout the research process to perform computation for statistical analysis.

Table 4. Persona of General Practitioner.

Name: Dr. Frank	**Status and Trust Level:** High
Role: General Practitioner (GP)	**Gender:** Male

Knowledge, Skills and Abilities: Dr. Frank knows how to treat patient and apply updated knowledge to improve quality of care.

Goals, Motives and Concerns: He wants to identify and prioritize his clinical practice performance improvement goals, and to track his progress towards those goals over time. Therefore, he measures and benchmarks his clinical practice with respect to the average performance of GPs in his area.

Usage Patterns: Dr. Frank is not a frequent user of Emnet. He only uses it once in a while.

Table 5. Persona of Public Health worker.

Name: Dr. Robin **Status and Trust Level:** High

Role: Clinical Researcher **Gender:** Male

Knowledge, Skills and Abilities: Dr. Robin knows how to design clinical research; statistical approaches used in clinical research and interpret analyses results; and economic and health outcome evaluation of health care interventions.

Goals, Motives and Concerns: To conduct research on the health and economic benefits and harms of different treatments, tests, procedures, and health care services for different groups of people.

Usage Patterns: Dr. Robin is a frequent user of *Emnet*.

4.3. Functional requirements

Based on the meetings with the representative users, personas and related works we extracted system requirements. We described the system boundary using event list technique (122). Then, the functional requirements are defined in the form of use cases and non-functional requirements are defined as atomic requirements ("The system shall...").

This section describes the functional requirements of *Emnet*. As the main goal of the thesis strongly ties to privacy, every functional requirement is designed to be in line with the concept of preserving privacy. Furthermore, privacy is described as a non-functional requirement in Section 4.4.

4.3.1. Event List and Use cases

In order to have a better understanding of the functional requirements, we make list of events in *Emnet*, corresponding inputs/outputs and description of each event. Then, the event list is further divided in to two sub-groups based on the two main tasks (data preparation and statistical computation) *Emnet* should perform. Table 6 and Table 7 show the event lists of data preparation and statistical computations respectively.

In this thesis, the following two phrases are used to describe two types of health institutions:

- "***set of health institutions***" refers to health institutions that are selected to participate in the research.
- "***participating data sources***" refers to health institutions that contributed data to the research.

Table 6. Event list in data preparation

#	Event Name	Input (IN)/output (OUT)	Summary
1	Specify research criteria	Criteria to define the data set (IN)	Get list of inclusion and exclusion criteria that select a research data
2	Process query	Set of criteria and ***Set of health institutions*** (IN) Formal query broadcasted to a **Set of health institutions** (OUT)	Use the set of criteria to form a query and broadcast it to the selected ***Set of health institutions***
3	Create data sets	Query string (IN) Stored data sets (OUT)	Execute the query and store the result locally at each ***Participating data source*** **
4	Display result of *Virtual Dataset*	Metadata of the created data sets (OUT)	Display the metadata of the *Virtual Dataset* to the user

*** The result of the query at a single data source is called a data set. The data sets across all* ***Participating data sources*** *collectively called a "Virtual Dataset"*

Table 7. Event list in statistical computations

#	Event Name	Input/output	Summary
1	Set Statistical Computation Inputs	Selected statistical function and variables (IN)	Get the selected statistical function and the variables to be computed by the function
2	Process query	Selected statistical function and set of variables (IN) Formal query broadcasted to ***Participating data sources*** (OUT)	Use the selected statistical function and set of variables to form a query and broadcast it to ***Participating data sources***
3	Perform statistical computation	Query string (IN) Computation results (OUT)	Execute the query against the *Virtual Dataset* and perform statistical computations
4	Display result of Statistical Computations	Statistical computation results (OUT)	Display the final result of the computation to the user

Based on the event list (see Table 6 and Table 7), we have created a UML use case diagram shown in Figure 12. The use case describes the user-system interaction in order to accomplish what the user requires to do with *Emnet*. We have continuously refactored it with new requirements and improvement of the existing one. The diagram represented six use cases. However, two of the use cases ("Process Query" and "Display Result") are used twice in

Emnet (during *Data Preparation* and *Statistical Computation*) as indicated by the «**include**» arrow.

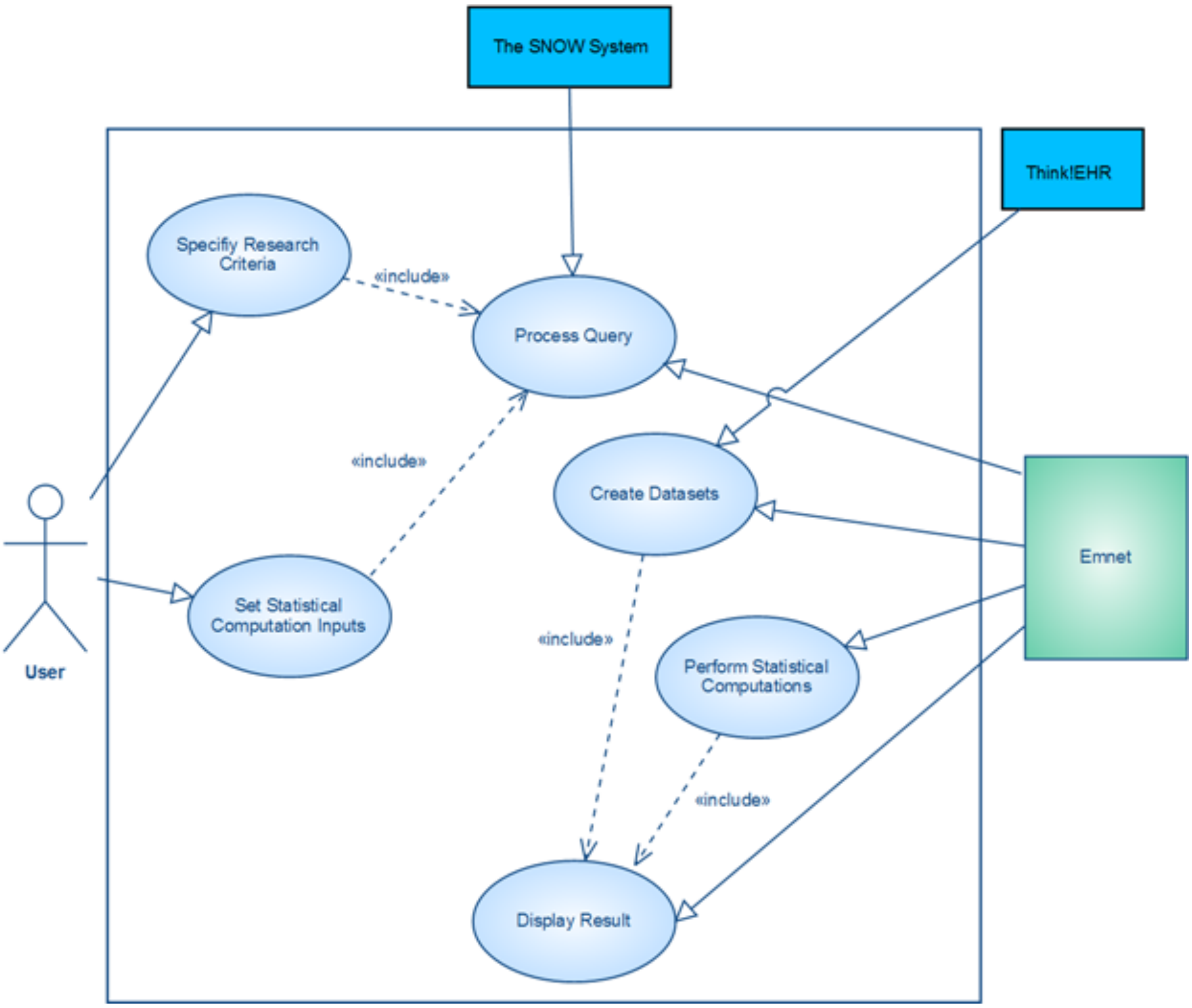


Figure 12. UML use case diagram.

An actor is a person or another system that interacts with a system. An actor can be *Active Actor* who is responsible for the task to be done or *Cooperative Actor* that helps the *Active Actor* to accomplish a task. Therefore, in this use case diagram, the *Active Actors* are:

- "***User***" – the end user of *Emnet* (e.g. epidemiologist or GP)
- "***Emnet***" - the proposed system

The detail description of functional requirements needed to accomplish each use case shown in Figure 12 is given below. The use cases are described in the order of execution. **Use case 2** and **Use case 4** are executed twice in *Emnet*. To express this, notation **A** and **B** are used in both use cases (i.e. **Use case 2A** and **Use case 4A** are executed in data preparation process and **Use case 2B** and **Use case 4B** are executed in statistical computation process).

Use case 1: Specify research criteria (Actor = User)

Purpose: To select the research data required for a research analyses

1. Select all inclusion criteria
2. Select all exclusion criteria
3. Select required attributes for analyses
4. Select ***set of health institutions***

Use case 2A: Process (data preparation) query (Actor = System)

Purpose: To build a formal query and deliver it to the ***set of health institutions***

1. Read the user input of inclusion and exclusion criteria
2. Build a formal query from the user input attributes
3. Read the user input about the selected ***set of health institutions***
4. Broadcast the query to the ***set of health institutions***

Use case 3: Create data sets (Actor = System)

Purpose: To prepare the research data

1. Run the query against the EHR at each health institution in ***set of health institutions***
2. Store the query results locally at the health institution

Use case 4A: Display (data preparation) results (Actor = System)

Purpose: To help the user visualize the created data set

1. Display metadata about the *Virtual Dataset* and descriptive statistics

Use case 5: Set statistical computation inputs (Actor = User)

Purpose: To prepare the needed data for the research analysis

1. Select statistical function
2. Select variables to be computed by the selected function

Use case 2B: Process (statistical computation) query (Actor = System)

Purpose: To build a formal analysis query and deliver it to ***participating data sources***

1. Read the user input for statistical function and variables
2. Build a formal query from the user inputs
3. Broadcast the query to ***participating data sources***

Use case 6: Perform statistical computation (Actor = System)

Purpose: To perform statistical computations on data sets created for the user

1. Perform privacy-preserving statistical computation on the *Virtual Dataset*
2. Store computation results
3. Return computation results

Use case 4B: Display (statistical computation) results (Actor = System)

Purpose: To make the user get the final result of the research analysis

1. Display the statistical computation results

4.4. Non-functional Requirements

Security and Privacy

As discussed in chapter 1 and 2, privacy is the main concern of this thesis. Consequently, it is the major non-functional requirement.

Thus, *Emnet* by no means moves patient identifiable data outside the health institution.

The system shall not allow any transfer of patient identifiable, personal data.

Privacy in this thesis also includes the privacy of the health institutions.

The system shall not display computation result of an individual health institution unless for the institution itself

The fact that patient's data - which is very sensitive information - is the core of this thesis, security is also the other major non-functional requirement. Thus, implementing strong security mechanisms plays a great role in preserving privacy in such a system. Security can be seen from three perspectives: authorization, authentication and access control.

The system shall only be used by authorized users.

The system shall not give access unless a user is authenticated.

The system shall permit or deny the use of a particular resource to a particular user based on the user credentials.

Appearance and Usability

The typical users of *Emnet* are non-technical users. If a system is quite new and complex to work with, users could easily get frustrated and might not be interested in using it. Consequently, system's appearance should be familiar to the users' experience. Besides, the functionalities of a system should be relevant and easily understandable by the users.

The system shall have the look and feel of statistical applications that the users are familiar with.

The system shall be easy to use to users who are familiar with other statistical software.

Extensibility

Implementation of all the functionalities, including statistical functions needed for research is beyond the scope of this thesis. However, *Emnet* should allow addition of any functionality provided that the functionality to be added does not violate the privacy requirement.

The system shall support the addition of extra feature to the system's functionality as long as the added feature does not violate privacy.

4.5. Summary

This chapter, in general, described the requirements specification. The major sources of the requirements were meetings with potential users and the literature review. Personas were also used as a participatory designed technique to compensate the lack of users' involvement in the requirement specification process.

Event list and use cases were used to explain the functional requirements. The main non-functional requirements of the thesis were also discussed in the last section of this chapter.

Chapter 5 Design

This chapter describes the general architecture and design of *Emnet*. It contains the brief explanation of the architecture, modules, interfaces, data and components of *Emnet* designed to fulfill the requirements specified in chapter 4.

5.1. Design Goals and Considerations

Privacy

The major design goal of *Emnet* is to develop a framework of a privacy-preserving statistical computation tool over distributed openEHR-based EHRs. Hereby, privacy remarkably affects both the architecture and design process. In other words, the interaction of the components in the prototype (the architecture) and how they are implemented and what functionalities they perform (the design) should comply with the privacy requirements specified in Chapter 4.

Appearance and usability

Appearance and usability of a system have the potential to make a system favorably used by the intended users. The user interface of the prototype should be easy to understand and use, even for people with no technical background. Thus, the appearance of the prototype should not be completely new to the users and the functionalities should not require technical knowledge to perform a given task.

Extensibility and modularity

Since all the necessary functionalities cannot be implemented in the scope of this thesis, the design process should consider extensibility of the functionality in the future. In addition, the health institutions that contribute data vary from research to research. Therefore, the design should enable to easily add or remove health institutions into *Emnet*. Together with extensibility, modularity is the important concept to consider during the design process. Modularity adds flexibility to components of *Emnet* when modification is needed. For example, if a new communication protocol is needed, it should be replaced without affecting the other components. Besides, modularity increases reusability of the components.

5.2. Architectural Design

Emnet is based on Service Oriented Architecture (SOA). The loose coupling feature of SOA fulfills the modularity and extensibility of our design considerations. The other reason for using this architectural pattern is security. It ensures that sensitive patient data used for computation should not leave its original heath institution. Any entity should only access the final result of computations. This enhances security and preserves privacy.

As described in chapter 4, *Emnet* should enable two main tasks: *data preparation* and *statistical computation*. These tasks are consecutive tasks. Thus, from now on we refer to them as two phases. In the first phase, a user prepares a research data by specifying a data preparation query that contains inclusion and exclusion criteria, and required attributes. In second phase, statistical computation will be performed on the prepared data. High-level description of how *Emnet* performs these tasks is given below.

1- Data preparation

One of the assumptions in *Emnet* is that the data are horizontally partitioned, where each health institution collects complete attributes of distinct set of individuals. Therefore, the user data preparation query will be broadcasted to a ***set of health institutions*** that will execute the query. Then, each health institution locally stores its data set of the data preparation query result. The data sets across the ***participating health institutions*** collectively make the research data, which we refer to it as *Virtual Dataset*.

2- Statistical computation

Following data preparation, the user can execute statistical computations on a *Virtual Dataset*. However, the statistical computation should be privacy preserving. We have used secure-multiparty computation techniques to enable joint privacy-preserving statistical computation between ***participating health institutions***. In section 5.3, we have described how privacy-preserving statistical computations can be performed on the *Virtual Dataset*.

In the following section the high-level architecture of *Emnet* and decomposition of its components are presented. Then, we described how these components are composed to perform the two phases described above.

5.2.1. Logical view - *Emnet* ‘s Components

As shown in Figure 13, based on SOA architecture, *Emnet* is organized in three layered components. The three major components of *Emnet* are *Client Application*, *Coordinating Agent*, and *Computing Agent*. The tasks executed by each component and its sub-components are briefly discussed below.

A. Client Application

Client application is the only component that is visible to the user; and it is connected to the *Coordinating Agent*. It provides interface for the user to perform tasks such as authentication, specifying a research inclusion and exclusion criteria, selecting a ***set of health institutions***, and selecting a statistical functions. This component also builds queries based on the user specification and makes the results of computations available to users. As shown in *Figure 13*, the *Client Application* has sub-components, such as *Communication* and *Display* modules.

While the *Display Module* handles the tasks on the user interface, the *Communication Module* accomplishes the message exchange between the *Client Application* and the *Coordinating Agent*.

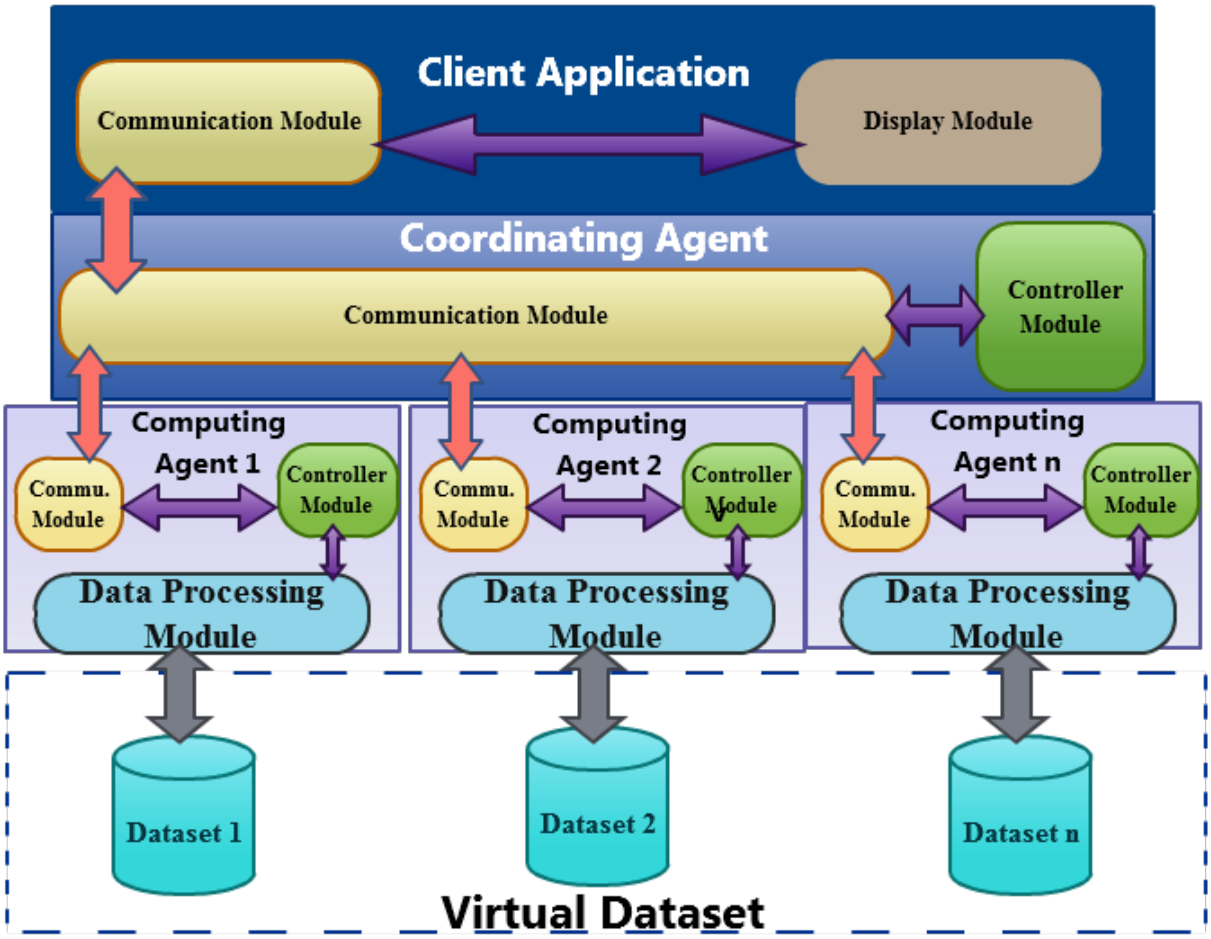


Figure 13. Modular system's components.

B. Coordinating Agent

As shown in *Figure 13*, the *Coordinating Agent* is a gateway for the *Client Application* to access remote health institutions that participate in a research. It broadcasts data preparation query received from the *Client Application* to remote health institutions; coordinate joint privacy-preserving computation of a user's statistical computation requests between health institutions; and return computation results to *Client Application.* The two sub-components of the *Coordinating Agent* are the *Communication Module* and the *Controller Module*. The *Communication Module* facilitates the communication of the *Coordinating Agent* with the other components in *Emnet*. However, the *Controller Module* broadcast data creation queries and coordinate joint privacy-preserving computations.

C. *Computing Agents*

The *Computing Agents* are the components located at each health institution. However, in Figure 13, only three *Computing Agents* are represented for brevity. Each *Computing Agent* executes data preparation query received from *Coordinating Agent* and locally store the resulting data set; and implements privacy-preserving protocols for joint computation with its data set. *Computing Agent* has three sub-components: *Communication Modul*e, *Controller Module* and *Data Processing* Module. The *Communication* and *Controller* Modules have the same responsibilities as the corresponding modules in the *Coordinating Agent*. However, the *Data Processing Module* executes data preparation query and locally store the query results.

5.2.2. Process view - *Emnet* 's System Structure

In this section we described how the *Client Application*, *Coordinating Agent*, and *Computing Agents* are composed to perform *Data Preparation* and *Statistical Computation* phases of *Emnet*. In both phases, as shown in Figure 14, the *Client Application* is connected to the *Coordinating Agent*, and the *Coordinating Agent* and the *Computing Agents* across health institutions are connected to each other. The detail design of each phase is explained in the succeeding subsections.

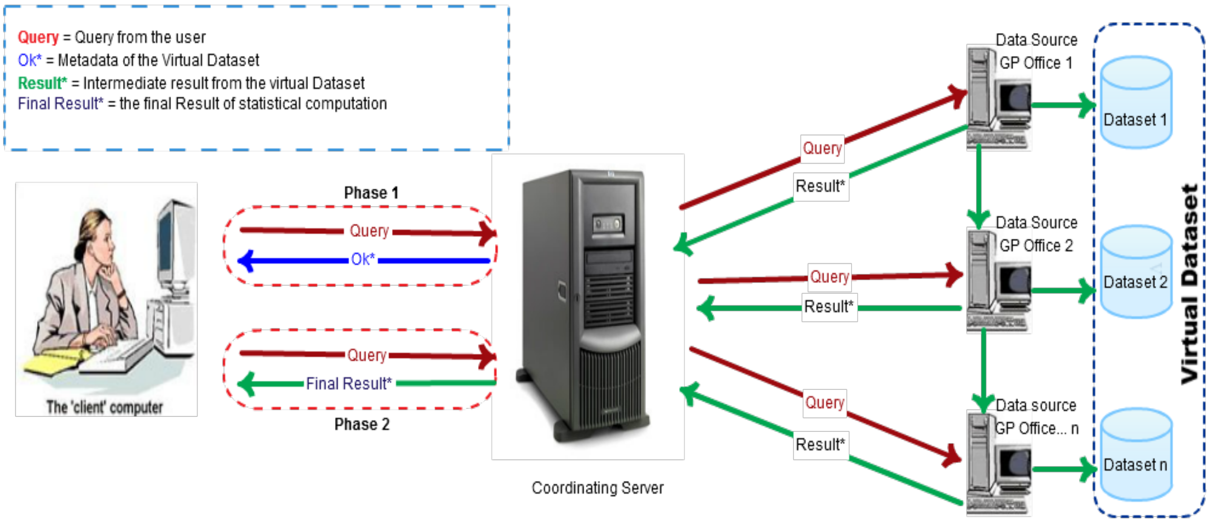


Figure 14. A two-phase process for privacy-preserving statistical query on distributed EHRs (123).

Phase I – Data Preparation phase

The *Data Preparation* phase creates a *Virtual Dataset,* which is a set of data set required to perform research analyses. The sequence diagram of the *Data Preparation* phase is shown in Figure 15. In this phase, the user sets the research inclusion and exclusion criteria and required attributes for a research via the *Client Application*. The *Client Application* uses the criteria to build an AQL query and sends it to the *Coordinating Agent*. Then, the

Coordinating Agent broadcasts to all *Computing Agents* that execute the AQL query against their EHRs. As discussed in the previous chapters, the EHRs across the health institutions are openEHR compliant. This enables to query all the EHRs using the same AQL query. Each health institution locally stores the result of the query (a data set) in a separate database. The data sets across all the institutions collectively are called *Virtual Dataset*. Therefore, the *Virtual Dataset* is the research data on which the research analyses are to be performed. Finally, the *Coordinating Agent* returns information about the *Virtual Dataset* as descriptive statistics results (note the descriptive statistics computation uses the techniques in the *Statistical Computation* phase) and metadata to the *Client Application.* Basically, the metadata is a tabular description of the created *Virtual Dataset*, which gives a visual description of the research data. As illustrated in *Figure 15,* phase I ends with displaying the descriptive statistics and metadata of the created *Virtual dataset* to the user.

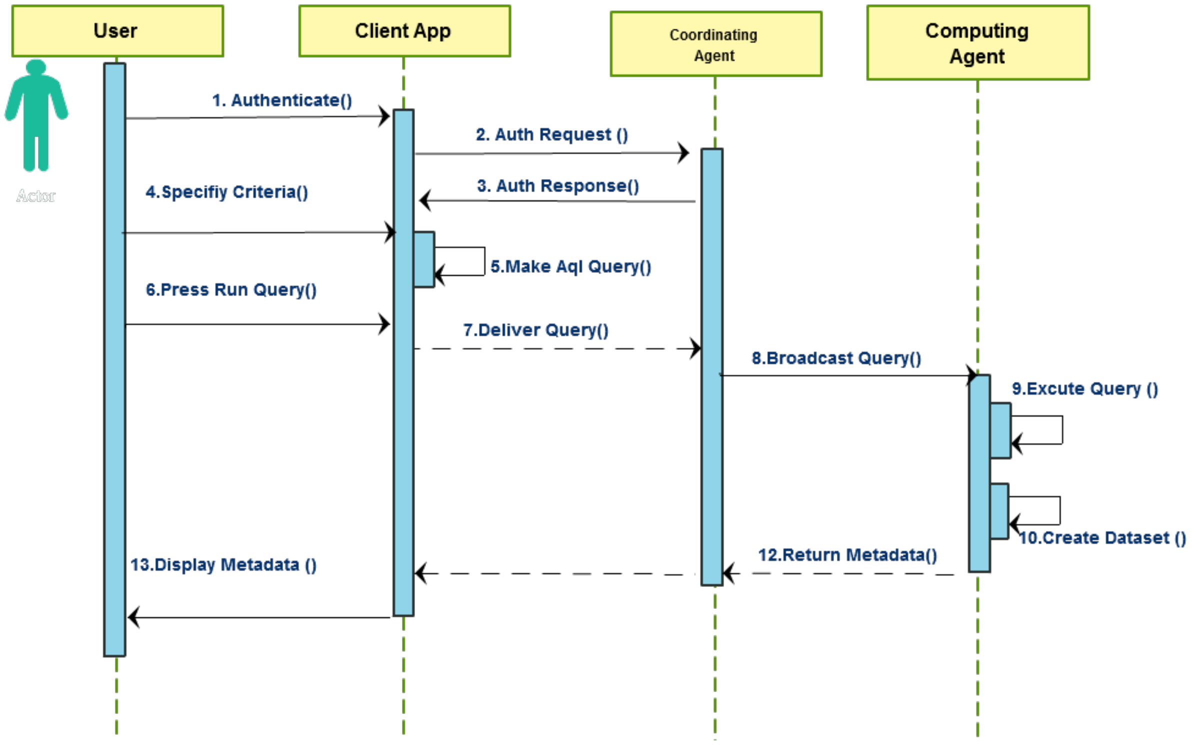


Figure 15. Sequence diagram of phase I.

Phase II – Statistical Computation

The *Statistical Computation* is the second phase that performs statistical computations on the *Virtual Dataset* created in phase I. As shown in *Figure 16*, phase II starts when the user selects the statistical function and the variables to be computed via *Client Application*. The *Client Application* creates a statistical computation query using the user inputs, and sends to the *Coordinating Agent*. The *Coordinating Agent* uses the computation graph and a secure summation protocol (see section 5.3) to jointly compute together with *Computing Agents* at the ***participating health institutions***. These processes are explained in further detail in section 5.3. Finally, the Coordinating Agent sends only the final, aggregated result of the computation to the *Client Application* where it is displayed to the user.

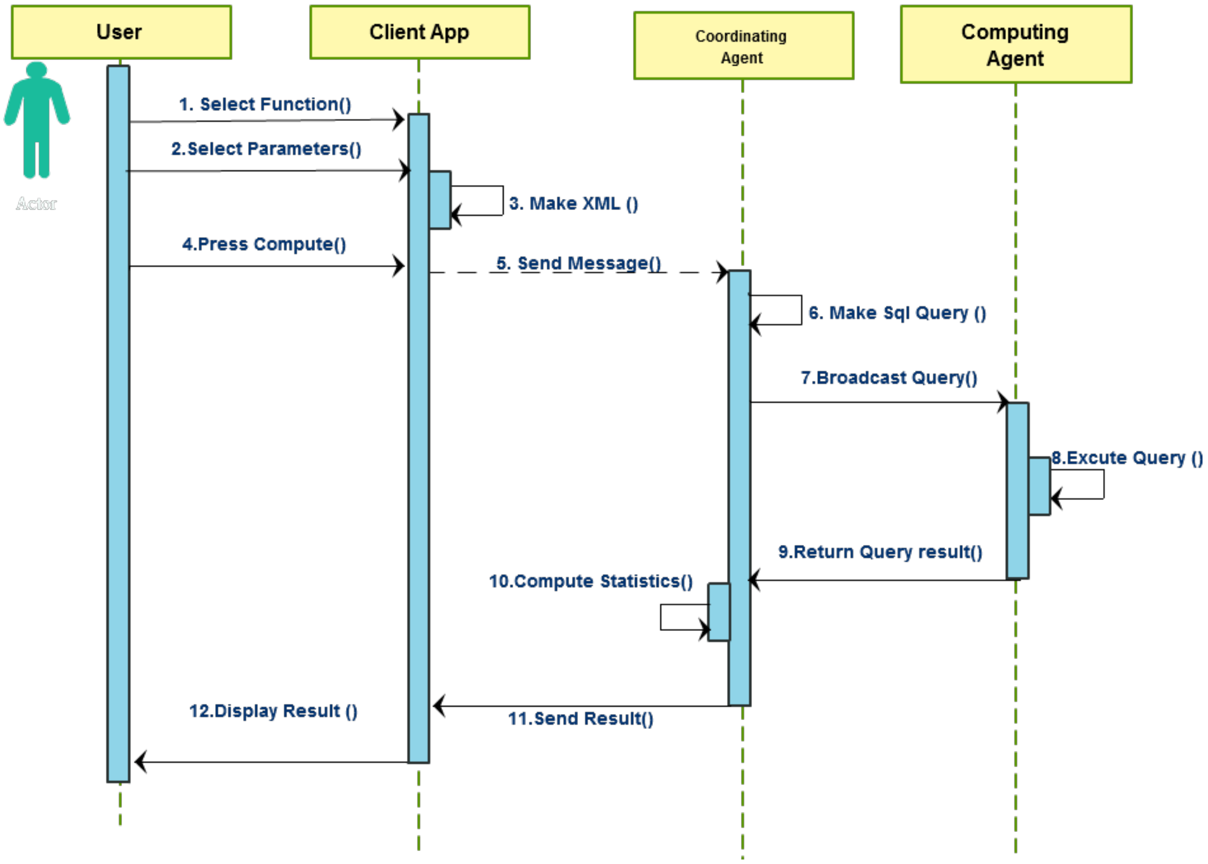


Figure 16. Sequence diagram of phase II.

5.3. Protocol Design

5.3.1. SMC Protocols Design

As discussed in chapter 2, SMC protocols are computing techniques for multiple parties to jointly compute a function over their inputs, and only computation results are revealed. The SMC protocols designed for the statistical computation described above are presented in this section.

Based on the assumption we made in section 1.3., the health institutions share their correct data and follow computation protocols. Therefore, protocols secure against semi-honest adversary is considered to be sufficient for computation among health institutions (76).

In this system, the basic building block is *Secure Summation* (see section 2.2.1). A statistical function is first decomposed into (expressed in) sub-computations of summation forms (see section 5.3.2). Then, each sub-computation is executed using a secure summation protocol.

As a result, private information of neither individual patient nor individual health institution is revealed. At the end of computation of the statistical function, only the final result and intermediate (sub-computation) results are revealed, which contain aggregate statistics of data from ***participating health institutions***.

As discussed in chapter 2, the choice of secure summation protocol is based on performance (such as communication overhead, and computation complexity) and privacy guarantee. As a result, in this thesis, we have chosen the SINE (103) protocol from the protocols reviewed in chapter 2 (see Table 2). The performance and the privacy guarantee of SINE are summarized as follow.

- It has communication overhead of b_p *(2n+1),* where b_p is bit information of the private value and *n* is the number of sites (***participating data sources***).
- It has computation complexity of O(n)
- It has strong privacy level assuming the C*oordinating Server,* that facilitates the computation process, is semi-trusted (follow the protocol and do not collude with any of the data sources to get additional information than what can be learned from computation results).

The nodes in SINE protocol are the health institutions in this system; and the coordinator in SINE protocol is the Coordinating agent in this system.

5.3.2. Decomposition of Basic Statistical Functions

A large number of linear and non-linear statistical functions can be decomposed into sub-computations of summation forms (124). Therefore, each sub-computation can be computed at each health institution and the results at the health institutions can be sum together using secure summation protocol. This makes sub-computations suitable to be parallelized (125–127) and as a result, statistical functions can be computed efficiently.

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

Summation (see equation 1a) of all patients' values of x across $\{H_1, H_2, H_3\}$ can be computed as follows. The summation can be expressed as equation 1b, where each institution locally sums its patients' values of x_i and then total summation will be summation of the local summation results of all institutions. The local summation result from individual institution contains aggregate data values of patients at the institution. Therefore, releasing it will not risk the patients' privacy. However, it can be considered private information of the institution.

Therefore, $\{H_1, H_2, H_3\}$ jointly execute a secure summation protocol (i.e. SINE) on their local summation results and only reveal the total summation result.

Function 1. Sum of X

$$\boldsymbol{sum}([x]) = \sum_{i=1}^{n}[x_i] \text{ .... (1a)}$$

$$\boldsymbol{sum}(x) = \sum_{i=1}^{j}[x_i] + \sum_{i=j+1}^{m}[x_i] + \sum_{i=m+1}^{n}[x_i] \text{................................. (1b)}$$

Similarly, any statistical function decomposed into sub-computations of summation forms can be computed in a privacy-preserving manner. Each sub-computation is computed in two computation rounds: each health institution locally computes the sub-computation; and the health institutions jointly execute SINE on their local results.

For a statistical computation request received from a user, the *Coordinating Agent* manages decomposition of the statistical function into sub-computations and coordinates execution of the sub-computations with the *Computing Agents* using the technique described above. *Computing Agents* execute local computations and jointly execute the secure summation protocol.

We have described decomposition of *count, mean, variance, covariance, standard deviation, Pearson's r* and *linear regression* into sub-computations of summation forms as follows.

Function 2. Count

$$\boldsymbol{count}() = n \text{ .... (2a)}$$

$$\boldsymbol{count}() = \boldsymbol{count}(1{:}j) + \boldsymbol{count}(j+1{:}m) + \boldsymbol{count}(m+1{:}n) \text{..............(2b)}$$

Where $\boldsymbol{count}()$ denotes the total counts across $\{H_1, H_2, H_3\}$; and $\boldsymbol{count}(1{:}j)$, $\boldsymbol{count}(j + 1{:}m)$, and $\boldsymbol{count}(m + 1{:}n)$ are counts at H_1, H_2, and H_3 respectively.

Count does not need decomposition, as it is already in summation form. Therefore, The *Coordinating Agent* sends count query to the *Computing Agents* and they count their patients' in the *Virtual Dataset*. Then, the *Coordinating Agent* and *Computing Agents* jointly execute secure summation of their local counts.

Function 3. Mean (Average) of X

$$\boldsymbol{mean}([x]) = \frac{(\sum_{i=1}^{n}[x_i])}{count()} \text{ ....(3)}$$

Equation 3 can be decomposed into sub-computations of summation and count. Thus, it can be expressed with equation (1a) and (2a).

$$\boldsymbol{mean}([x]) = \frac{\boldsymbol{sum}([x])}{\boldsymbol{count}()}$$

Using the technique described above, the *Coordinating Agent* coordinates execution of equation (1a) and (2a) with the *Computing Agents*, and use the results to execute $\boldsymbol{mean}([x])$. The rest of the decomposed statistical functions will be computed in the same manner.

Function 4. Variance of X

$$\boldsymbol{var}([x]) = \frac{(\sum_{i=1}^{n}([x_i]-mean([x]))^2)}{count()} \text{..}(4)$$

As shown in equation (5), we denote $([x_i] - \boldsymbol{mean}([x]))$ at x_i as $\boldsymbol{Interm}([x_i])$.

$$\boldsymbol{Interm}([x_i]) = \left([x_i] - \boldsymbol{mean}([x])\right) \text{... (5)}$$

Then, variance can be expressed with equation (4) and (5),

$$\boldsymbol{var}([x]) = \frac{(\sum_{i=1}^{n} \boldsymbol{Interm}([x_i])^{\mathbf{2}})}{\boldsymbol{count}()}$$

Function 5. Standard Deviation of X

$$\boldsymbol{sdev}([x]) = \sqrt{\boldsymbol{var}([x])} \text{ .... (6)}$$

Since variance of x already computed previously, *equation 5*, it is easy to compute standard deviation of *x*.

Function 6. Covariance of X and Y

$$\boldsymbol{Covar}([x][y]) = \frac{\sum_{i=1}^{n}([x_i]-mean([x]))([y_i]-mean([y]))}{count()} \text{ ....(7)}$$

Similar to equation (5) we can compute the following two intermediate values at x_i and y_i

$$\boldsymbol{Interm}([y_i]) = [y_i] - \boldsymbol{mean}([y]) \text{ .... (8)}$$

$$\boldsymbol{Interm}([x][y]) = ([x_i] - \boldsymbol{mean}([x]))([y_i] - \boldsymbol{mean}([y])) \text{.........................(9)}$$

Equation (8) and (5) can be substituted into equation (9) and $\boldsymbol{Interm}([x][y])$ can be expressed as,

$$\boldsymbol{Interm}([x][y]) = \boldsymbol{Interm}([x_i]) * \boldsymbol{Interm}([y_i])$$

Substitute equation (9) into equation (7) and covariance can be expressed as,

$$\boldsymbol{Covar}([x][y]) = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{Interm}([x][y])}{\boldsymbol{count}()}$$

Function 7. Pearson's R (Correlation) of X and Y

$$\boldsymbol{r}([x][y]) = \frac{\sum_{i=1}^{n}([x_i]-\boldsymbol{mean}([x]))([y_i]-\boldsymbol{mean}([y]))}{\sqrt{\sum_{i=1}^{n}([x_i]-\boldsymbol{mean}([x]))^2 \sum_{i=1}^{n}([y_i]-\boldsymbol{mean}([y]))^2}} \text{ ……... (10)}$$

Substitute equation (6) and (7) into equation (10), and Pearson's r (correlation) of x and y can be expressed as,

$$\boldsymbol{r}([x][y]) = \frac{\boldsymbol{Covar}([x][y])}{\sqrt{\boldsymbol{sdev}([x])\boldsymbol{sdev}([y])}}$$

Function 8. Simple Linear Regression of X and Y

$$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}.\boldsymbol{x} \text{ ………………….. (11)}$$

Where, $\boldsymbol{y}$ = dependent variable, $\boldsymbol{x}$ = independent variable, $\boldsymbol{\beta}_0$= intercept, and $\boldsymbol{\beta}_1$ = slope

$\boldsymbol{\beta_1}$ is computed as,

$$\boldsymbol{\beta_1} = \frac{\sum_{i=1}^{n}([x_i]-mean([x]))([y_i]-mean([y]))}{\sum_{i=1}^{n}([x_i]-mean([x]))^2} \text{……………………………….(12)}$$

Substitute equation (4) and (6) into the above equation, and $\boldsymbol{\beta}_1$ can be expressed as,

$$\boldsymbol{\beta_1} = \frac{\boldsymbol{Covar}([\boldsymbol{x}][\boldsymbol{y}])}{\boldsymbol{var}([\boldsymbol{x}])}$$

And $\boldsymbol{\beta_0}$ can be expressed using $\boldsymbol{mean}([x])$, $\boldsymbol{mean}([y])$, and $\boldsymbol{\beta}_1$ as,

$$\boldsymbol{\beta_0} = \boldsymbol{mean}([y]) - \boldsymbol{\beta_1}\,.\boldsymbol{mean}([x]) \text{………………………………………. (13)}$$

The statistical functions described above are decomposed into sub-computations of summation forms. Then, each sub computations are executed using the steps shown for *count* and *sum* using SINE protocol. A table showing decomposition of more statistical functions in summation forms is attached in Appendix B.

5.3.3. Computational Graph

In the previous sub-section we described how statistical functions can be decomposed into sub-computations of summation forms, and how these sub-computations are computed among the *Coordinating Agent* and *Computing Agents* using secure summation protocol (SINE).

Figure 17 demonstrates an abstract computation graph that shows the dependencies between the statistical functions implemented in this system. In the computation graph, the nodes represent statistical functions and the edges point to the direction of dependency between nodes. The dependency indicates the statistical function found in the higher level depends on the result of the functions found in the lower level. For example, if a user selected *Pearson's* *r,* the lower level functions, such as *Covariance, Standard Deviation, Mean* and *Sum* should be computed hierarchically.

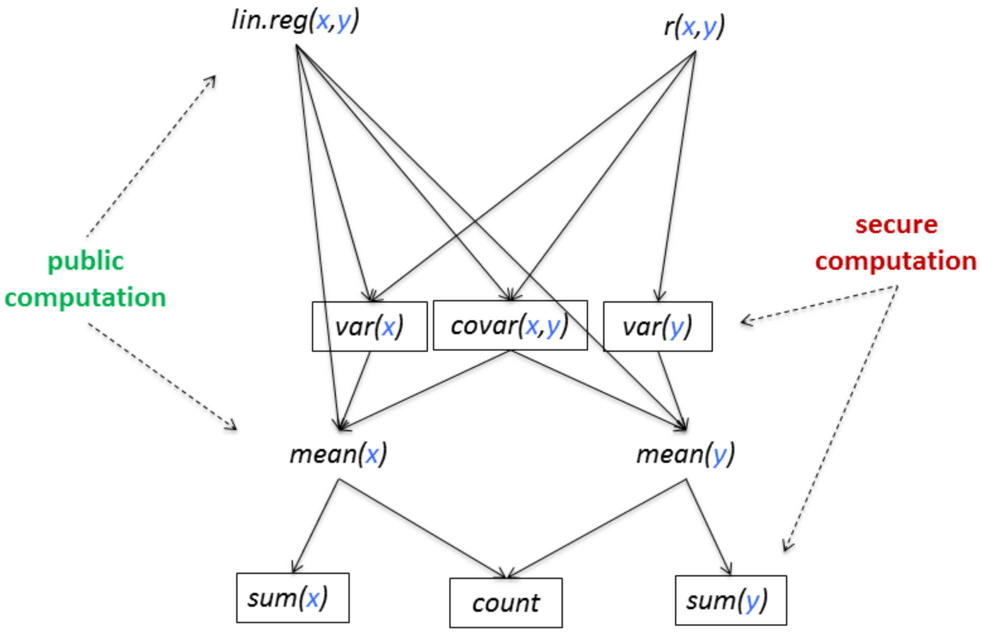


Figure 17. Computational graph of the decomposed statistical functions.

As shown in Figure 17, the computation graph has secure and public type of computations. The functions in a box should be securely computed using the technique described in section 5.3.2; because they are directly computed on individuals' data at the *Computing Agents*. Whereas, the functions outside a box can be computed at the *Coordinating Agent* since they are based on only the results of lower level functions. The final computation results of the nodes (functions) are considered as non-sensitive information; because they are the aggregated result of all the data sources.

5.4. Data Design

The original data used by *Emnet* is horizontal partitioned General Practitioners offices EHR data. Normally, *Emnet* assumes that all the EHR data involved in the computation are in openEHR format, which is not currently the case in the practice. However, the current strategic plan of Norwegian Health authorities is currently encouraging EHR vendors to adopt openEHR to enable health information interoperability (23). For this thesis, we created openEHR data to be used as data sources. The openEHR data preparation process is broadly described in the chapter 7.

During *Data Preparation* phase, every data source executes the *Data Preparation* query and locally store the result in a separate relational database. The main reasons for having separate database are described below:

- To prevent modification of the EHR during *statistical query processing*. The *statistical computation* phase produces intermediate computation results and additional database columns are created to temporarily store these results.
- To prevent multiple access to the EHR. Statistical computations involve several computations on a research data. Therefore, storage of a research data in a separate database prevents frequent execution of queries against the EHR.
- To facilitate data cleaning for the research. In order to modify the quality of the EHR data as it fits for the research's need, data cleaning and transformation should be applied. Hence, having separate data set avoids the possible alteration of the original EHR data.

5.5. Interface Design

To fulfill the ***Appearance and Usability*** non-functional requirement, the design of the user interface is made to be simple and as easy to use as possible. Besides, most of the interface components are familiar to the user. The first interface the user gets when typing the URL (address of the web application) of the *Client Application* is the *Login Interface*. As shown in *Figure 18,* the *Login Interface* is designed in a simple and familiar way. After successful authentication, the main interface will be available. It has *Data Preparation* and *Statistical Computation* interfaces contained in two tabs. The role and usage of these interfaces are described below.

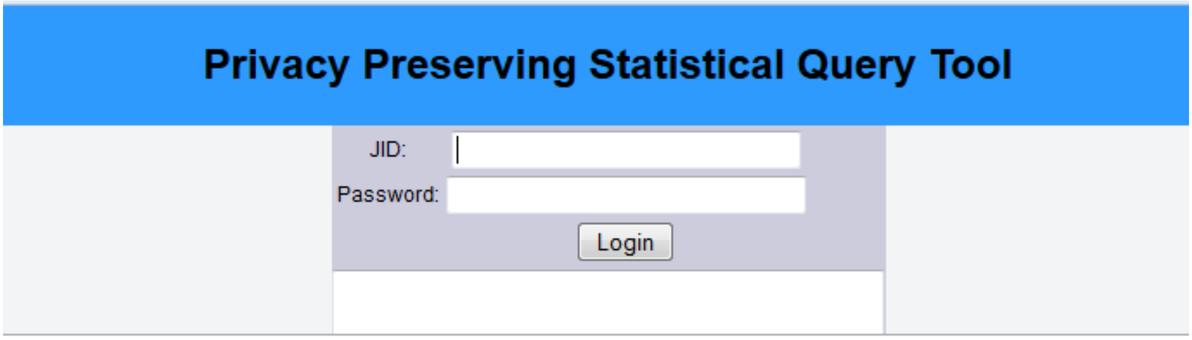


Figure 18. Login page of the Client Application.

A. ***Data Preparation*** – It is the first interface that is displayed following successful login and it is named as "***Create Dataset***". It enables the user to specify criteria for *Data Preparation* query. As shown in Figure 19, the interface has the list of medical concepts arranged in tree view (indicated by arrow 1) so that the user can drag and drop into the *drop area* indicated by arrow 2. The *drop area* is the working area that holds the selected research criteria used to build an AQL query. The *Add Criteria* button, shown by arrow 3, helps to add more *drop areas* that hold new criteria. The *Data Preparation* interface also has the button, shown by arrow 4, to build and submit the query to the *Coordinating Agent*. As indicated by arrow 5, the interface has a display area where the status of the *Data Preparation* process is notified to the user.

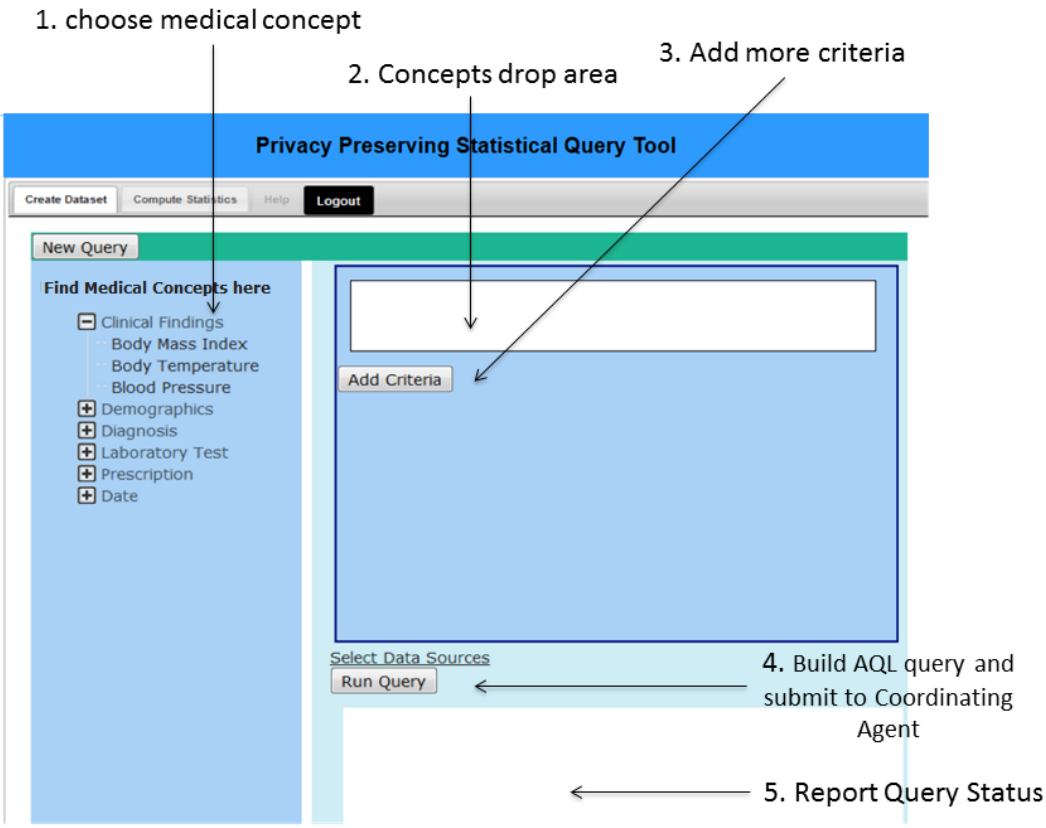
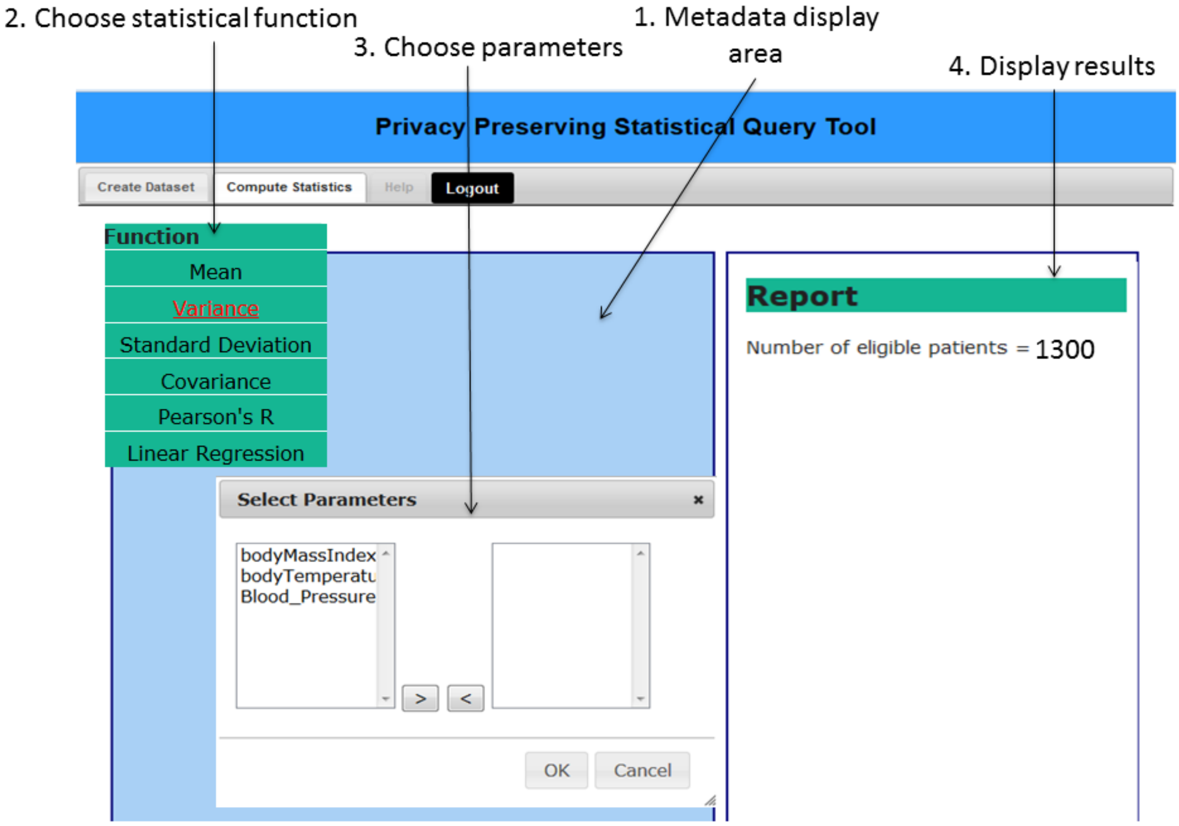


Figure 19. The Data Preparation Interface of the Client Application.

B. ***Statistical Computation*** – It is the interface that enables the user to perform statistical computations on the created *Virtual Dataset*. To make it user-familiar, the design of the *Statistical Computation* interface is inspired by SPSS[7] (a commonly known statistical analysis tool). As depicted in Figure 20, this interface has four main parts. Arrow 1 indicates the area that displays a tabular description of metadata of the created *Virtual Dataset*. Arrow 2 and 3 illustrate the list of statistical functions and the list of variables arranged in the dialog box respectively. These two are the inputs specified by the user for *Statistical Computation* query. As indicated by arrow 4, the final results of the statistical computations are displayed in the report section of the interface.



Number of eligible patients = 1300

Figure 20. The Statistical Computation Interface of the Client Application.

5.6. Summary

The chapter covered a wide range of ideas and concepts that constitute the design of *Emnet*. The first section described the factors taken into consideration during the design process. Next, the general overview of the architectural design is described, such as system architecture with regard to logical and process views, major components of *Emnet* and how they interact to each other to accomplish tasks requested by the user.

The chapter then illustrated the SMC protocol used in the design along with the rationale for choosing the specific secure sum protocol (SINE protocol). Further down, the chapter presented decomposition of the statistical functions used in epidemiological research into sub-computations of summation forms. It also illustrated the computation graph made of the decomposed statistical functions. The next section described the data design process and the structure of the data used in *Emnet*. The chapter finally concluded by discussing the user interface design.

Chapter 6 Implementation

This chapter discusses implementation details of the distributed privacy-preserving statistical computing on EHRs designed in chapter 5. The first section briefly explains programming languages and technologies used for the implementation. The rest of the sections present implementation details of each component of *Emnet*.

6.1. Programming Languages and Technologies

As *Emnet* is composed of different components responsible for different tasks, the programming languages and technologies used for implementation also differ. The following subsections describe the programming technologies used to implement the *Client Application*, *Coordinating Agent* and *Computing Agent*, and communication between these components.

6.1.1. Communication Technology

As discussed in chapter 5, *Emnet* uses service oriented architecture (SOA) architectural pattern, which is mainly characterized by services provided among the different independent components of *Emnet* through message exchange. In other words, one component sends a message to another component to get a service or to get a work done. To enable communication between the components of *Emnet*, XMPP messaging technology is used in our implementation.

XMPP is a set of application protocol created for real-time message communication and presence. The reason for selecting XMPP in this thesis is described as follows:

- *Emnet* is a part of the Snow system, which uses XMPP for communication between its components.
- The XMPP protocols are open and easy to use; besides, there are multiple implementations of clients, servers, and libraries that can be customized for use.
- All health institutions (i.e. GPs and hospitals) in Norway are connected via Norwegian Health Network (HealthNet), which is aimed to enable secure electronic communication between the institutions[9]. The Norwegian Code of Conduct for secure health information exchange (128) requires that all communication requests should be initiated from inside a health institution. It was easier to achieve this requirement using XMPP.

XMPP technology is based on client/server architecture, where clients are interconnected through relaying servers. We have used Openfire XMPP server. Openfire

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

and *Coordinating Agent* are installed in the HealthNet. We implemented each component of *Emnet* (*Client Application*, *Coordinating agent*, and *Computing Agent)* as an XMPP client identified by Jabber Id (JID). The clients authenticate and connect to the Openfire server, and the connections last long. Messages are sent between *Emnet*'s components (clients) using their JID. Therefore, messages are sent to a *Computing Agent* through connection it has initiated from inside the health institution.

- XMPP has strong security features such as the possibility of isolating XMPP server from public network like Internet, the incorporation of SASL (Simple Authentication and Security Layer) and TLS (Transport Layer Security) in the core of XMPP specifications and the end- to-end encryption mechanism.

6.1.2. Programming Language and Technologies used for the Components

Figure 21 shows the overview of the components and sub-components of the whole system together with the programming languages and tools (written in blue texts) used in each component. To simplify the demonstration, only one *Computing Agent* is included in the architecture shown in the figure; however in reality a *Computing Agent* is installed at each *health institution*.

The *Client Application* is a web-based application. It is implemented in JavaScript, HTML5 and CSS3. The *Communication Module* is implemented using a JavaScript library called Strophe. Strophe enables a web based XMPP messaging. Normally, web browsers support HTTP protocol and they don't have built-in support for XMPP protocol. Therefore, we used a connection manager called BOSH (129) that tunnels XMPP sessions over HTTP connections efficiently. We used the built-in BOSH implementation in Openfire server to handle the connection and communication of *Client Application* and the Openfire server.

We have implemented the *Coordinating Agent* and *Computing Agent* in Java. The main rationale for choosing Java is because it is a platform independent language. It gives the flexibility of implementing our system at several health institutions without the need to require them to have a specific type of operating system. Another rationale for using java was to easily use the Think!EHR interface, which is implemented in Java. Moreover, in the future it will be easier to make the components as Openfire plugins, to integrate it to the Snow system.

The *Communication Module* of the *Coordinating Agent* and *Computing Agent* are implemented XMPP messaging functionalities using a Java library called SMACK (119). SMACK is an Open Source XMPP library.

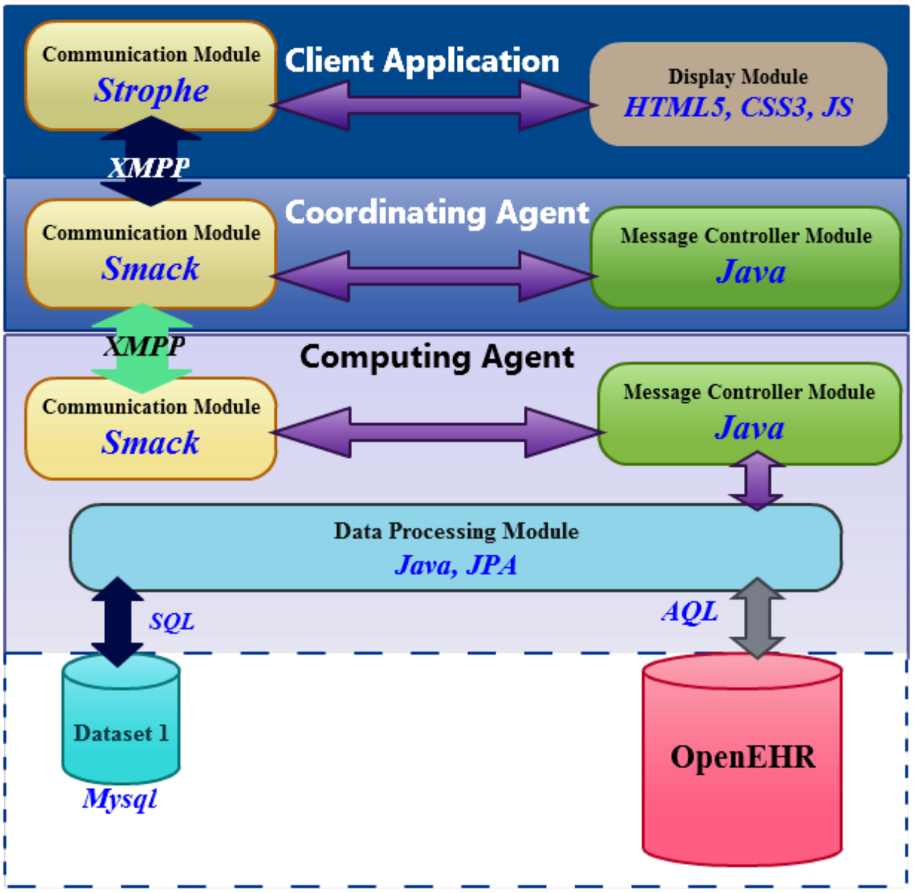


Figure 21. Technologies used to implement the individual components of the prototype.

Since the EHRs are OpenEHR-based (Think!EHR core in our case), the *Data Preparation* queries are built in AQL in the first phase of *Emnet*. The *Data Processing* module of the *Computing Agent* use Think!EHR interface and library to execute AQL queries against the Think!EHR core. It uses JPA object/relational mapping technology to store the query results in MySQL relational database. JPA is also used to run SQL queries against MySQL.

6.1.3. Communication Protocol

XMPP protocol communicates using XML message. In addition, we have designed an XML message protocol that defines *Data Preparation and Statistical Computation* queries and responses. The XML message is sent inside XMPP XML message stanza. The XML message has the following elements:

QueryId – a unique id used to identify messages that belong to the same communication session.

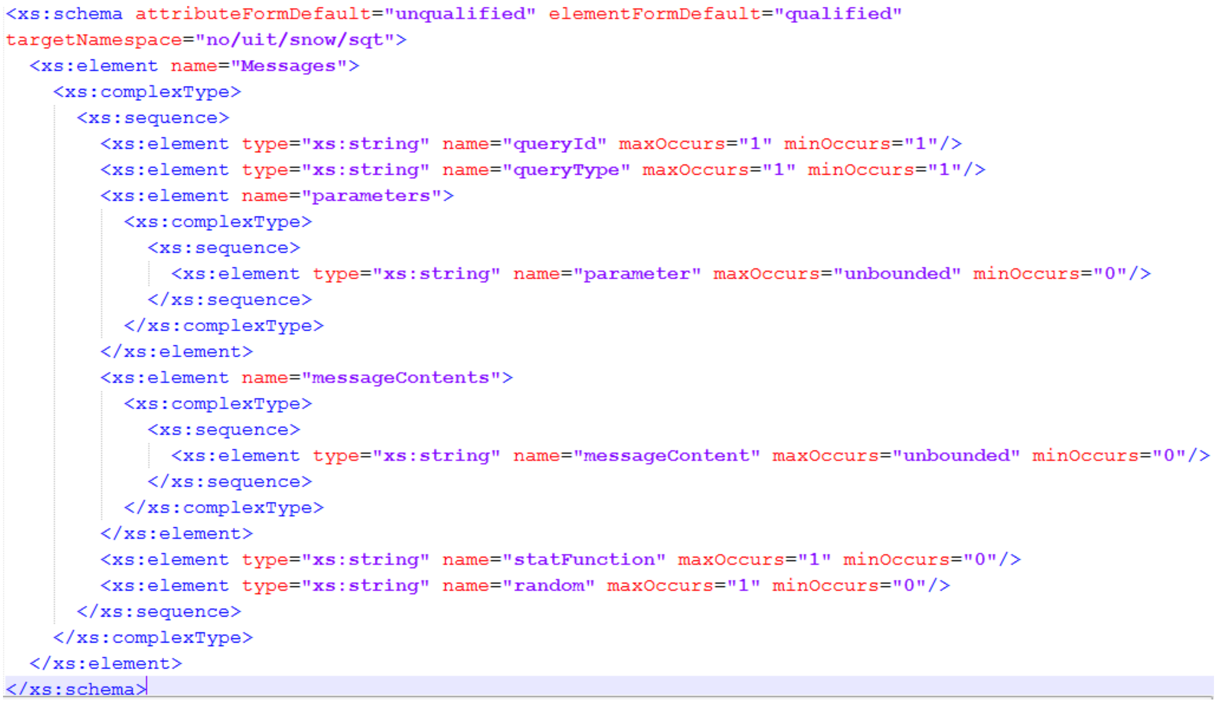
QueryType – identifies whether a message is either *Data Preparation,* or *Statistical Computation* query, or response.

Parameters – holds parameter(s) required for *Statistical Computation*.

MessegeContents – holds a query string to be executed against the database or computations results.

StatFunction – holds a statistical function selected by the user.

Random – holds a random number used in the SINE protocol.



```
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
targetNamespace="no/uit/snow/sqt">
  <xs:element name="Messages">
    <xs:complexType>
      <xs:sequence>
        <xs:element type="xs:string" name="queryId" maxOccurs="1" minOccurs="1"/>
        <xs:element type="xs:string" name="queryType" maxOccurs="1" minOccurs="1"/>
        <xs:element name="parameters">
          <xs:complexType>
            <xs:sequence>
              <xs:element type="xs:string" name="parameter" maxOccurs="unbounded" minOccurs="0"/>
            </xs:sequence>
          </xs:complexType>
        </xs:element>
        <xs:element name="messageContents">
          <xs:complexType>
            <xs:sequence>
              <xs:element type="xs:string" name="messageContent" maxOccurs="unbounded" minOccurs="0"/>
            </xs:sequence>
          </xs:complexType>
        </xs:element>
        <xs:element type="xs:string" name="statFunction" maxOccurs="1" minOccurs="0"/>
        <xs:element type="xs:string" name="random" maxOccurs="1" minOccurs="0"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

Figure 22. Communication protocol XML schema.

Figure 22 shows the XML schema of the communication protocol defined in XSD (XML Schema Definition language). However, based on the type of a message, some of the XML elements may not be used (indicated by *minOccurs="0"* in Figure 22). For example, the ***StatFunction*** element is not used in the *Data Preparation* queries.

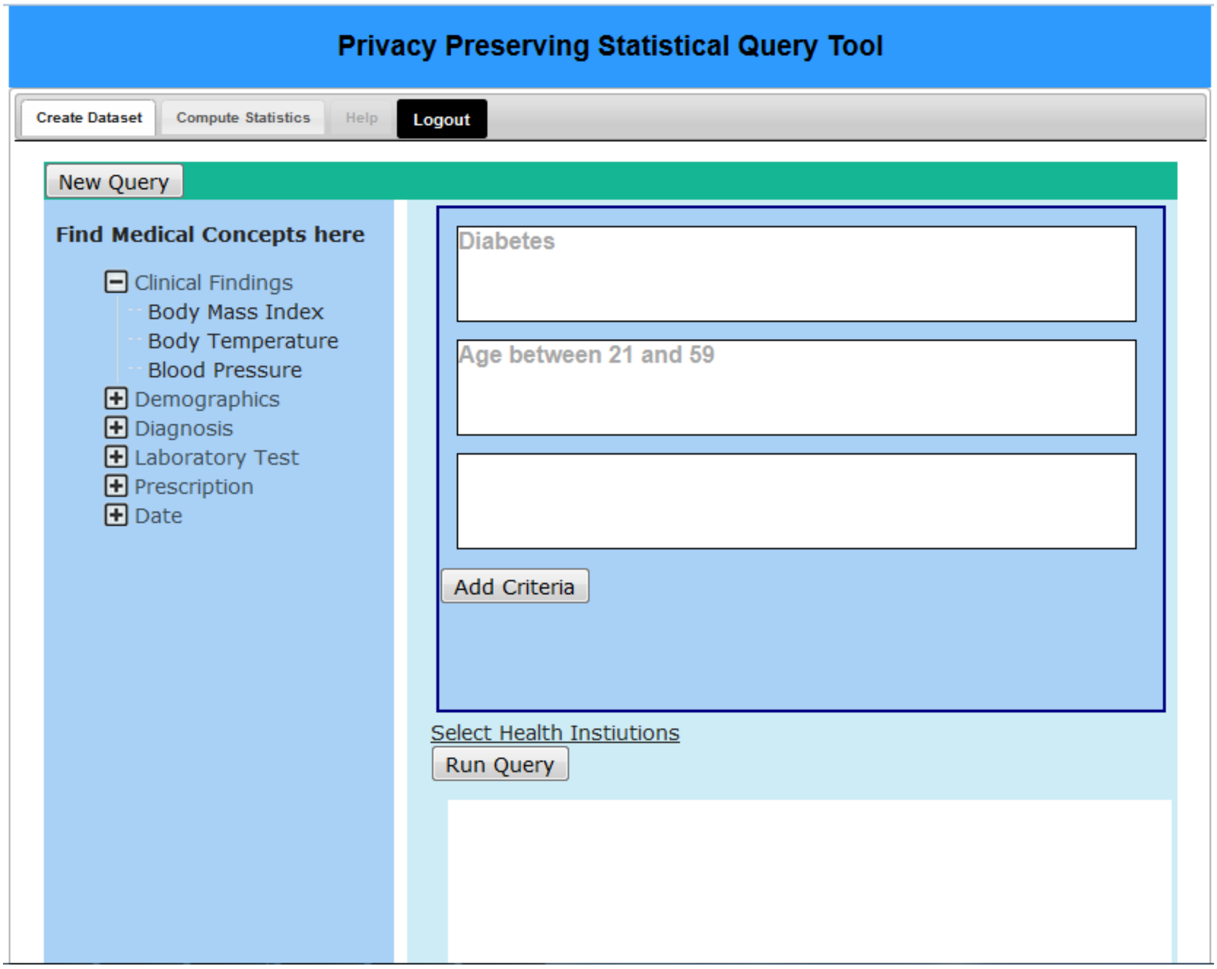
6.2. Client Application

The Client Application is implemented as a web application, because:

- Web applications only need a web browser on the user machine. As a reuslt, there is no need of installing any kind of software besides a browser.
- The application can easily be updated without users to update.
- Authorized users can easily acess the application from anywhere over the internet.

The application has a simple login interface where the user inserts a Jabber ID (JID) and password and the Openfire server authenticates. The Openfire server has its own database that store JIDs and passwords of registered users.

After successful authentication, the user sees the *Create Dataset* page (see Figure 23). This is the place where the user specifies the research *Data Preparation* query. The query specification procedure starts with dragging *medical concepts* to the *drop area*. The user can add concepts by pressing the "*Add Criteria*" button, which creates additional *drop area*. For each additional concept, the user can choose whether it is Inclusion or Exclusion. Moreover, when a concept is dropped on the *drop area*, a dialog box appears to specify constraints (e.g. if a user dropped *Body Mass Index* onto the *drop area*, the application asks the user to set constraints of the value to be selected).



Diabetes

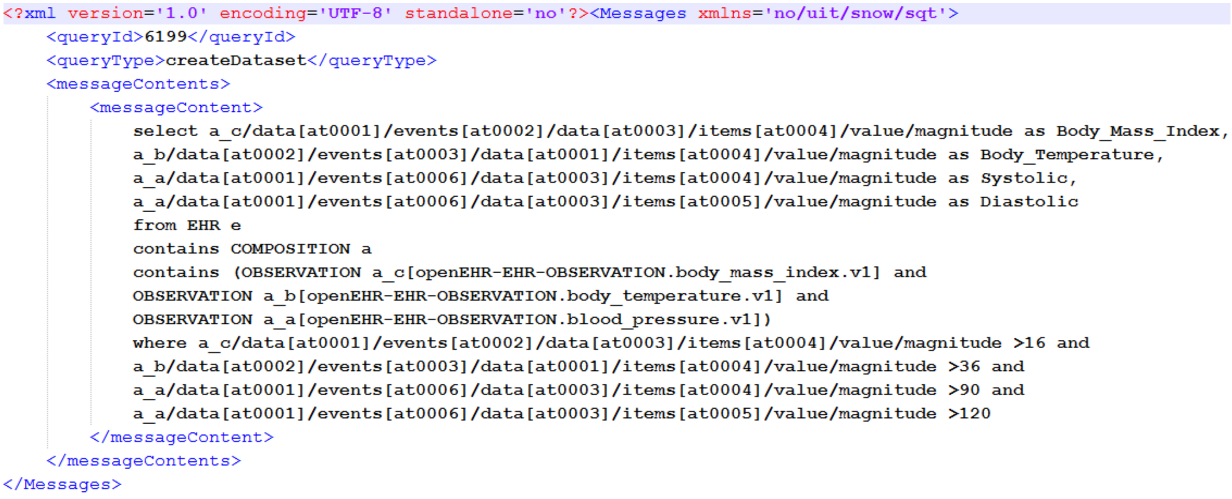
Age between 21 and 59

Figure 23. Research Criteria specification on the Client Application.

Based on some criteria (for e.g. geographical area), the user can choose a set of health institutions to use as data sources by clicking "Select Health Institutions" button. This is to identify the set of health institutions (data sources) that will contribute data for the research.

Finally, following the user presses the “Run Query” button; the *Client Application* builds an AQL query from the user inputs. To build the query, the application uses the archetypes that correspond to the medical concepts (criteria) selected by the user interface. The archetypes are collected from the Norwegian archetype registry available in the Clinical Knowledge Manager (CKM) registry[10].

The *AQL query string* together with generated random *queryId* and *queryType* forms an XML message. Snippet 1 shows an example XML message that contains *queryId* of ***6199***, *queryType* of ***CreateDataset*** and *MessageContent* of 13 lines of ***AQL query***.



```
<?xml version='1.0' encoding='UTF-8' standalone='no'?><Messages xmlns='no/uit/snow/sqt'>
    <queryId>6199</queryId>
    <queryType>createDataset</queryType>
    <messageContents>
        <messageContent>
            select a_c/data[at0001]/events[at0002]/data[at0003]/items[at0004]/value/magnitude as Body_Mass_Index,
            a_b/data[at0002]/events[at0003]/data[at0001]/items[at0004]/value/magnitude as Body_Temperature,
            a_a/data[at0001]/events[at0006]/data[at0003]/items[at0004]/value/magnitude as Systolic,
            a_a/data[at0001]/events[at0006]/data[at0003]/items[at0005]/value/magnitude as Diastolic
            from EHR e
            contains COMPOSITION a
            contains (OBSERVATION a_c[openEHR-EHR-OBSERVATION.body_mass_index.v1] and
            OBSERVATION a_b[openEHR-EHR-OBSERVATION.body_temperature.v1] and
            OBSERVATION a_a[openEHR-EHR-OBSERVATION.blood_pressure.v1])
            where a_c/data[at0001]/events[at0002]/data[at0003]/items[at0004]/value/magnitude >16 and
            a_b/data[at0002]/events[at0003]/data[at0001]/items[at0004]/value/magnitude >36 and
            a_a/data[at0001]/events[at0006]/data[at0003]/items[at0004]/value/magnitude >90 and
            a_a/data[at0001]/events[at0006]/data[at0003]/items[at0005]/value/magnitude >120
        </messageContent>
    </messageContents>
</Messages>
```

Snippet 1. Data Preparation XML message containing the AQL query built from the user-specified criteria.

The *Client Application* uses Strophe JavaScript library to send the XML message as XMPP message to the *Coordinating Agent* identified by its JID. The *Client Application* displays the *Data Preparation* query response from the *Coordinating Agent*, such as metadata of the *Virtual Dataset* created (tabular description of the stored database to help the user visualize the research data) and the number of patients who satisfy the specified criteria.

Following the *Virtual Dataset* creation, the user can perform *Statistical Computation*. In this phase, the user selects a statistical function of interest and the variables to be computed on by the selected function. Based on the input from the user, the *Client Application* creates a new XML message and sends as XMPP message to the *Coordinating Agent* identified by its JID. The *Client Application* also displays the result of the *Statistical Computation* query returned from the *Coordinating Agent*.

[10] http://arketyper.no

6.3. Coordinating Agent

The *Coordinating Agent* serves as a mediator between the *Client Application* and the *Computing Agents* (data sources); and it manages and coordinates execution of *Data Preparation* and *Statistical Computation* queries together with the *Computing Agents*. As shown in Figure 24, it has *Communication Module* and *Message Controller Module* sub-components. The *Communication Module* sends and receives messages from the Client Application and Computing Agents. The *Communication Module* has a list of the JIDs of all the *Computing Agent*s and the *Client Application.*

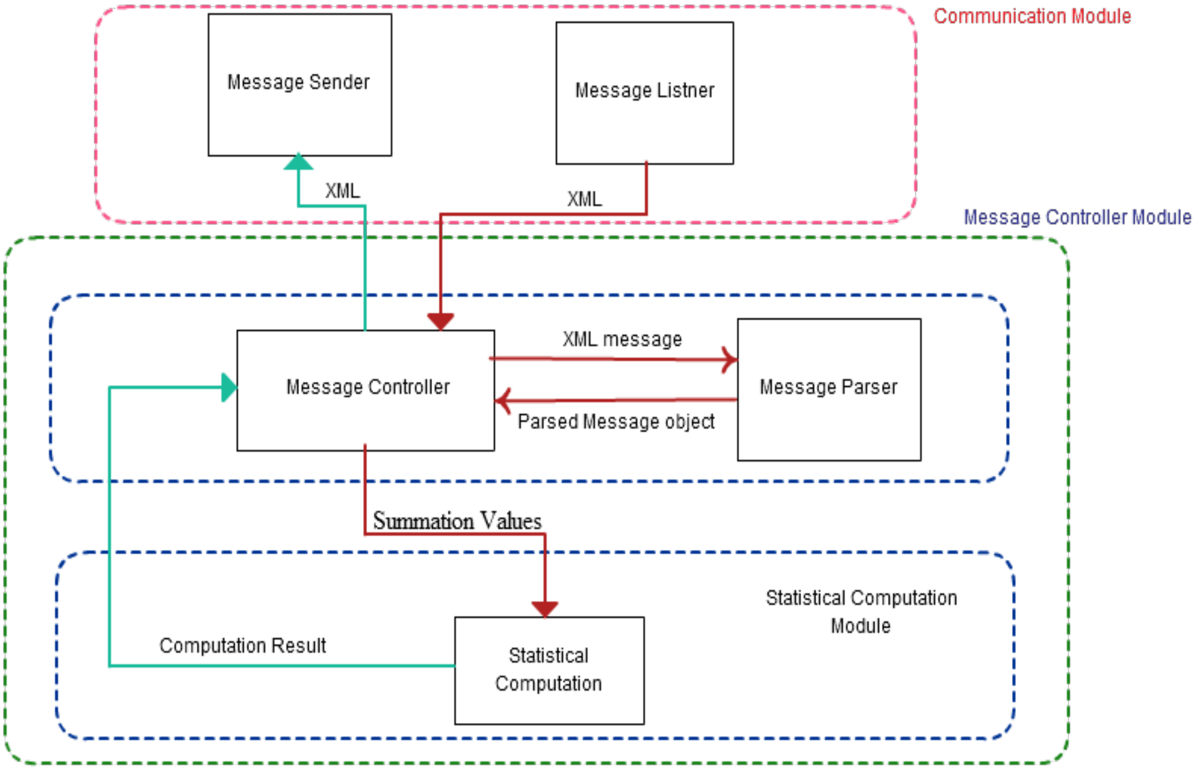
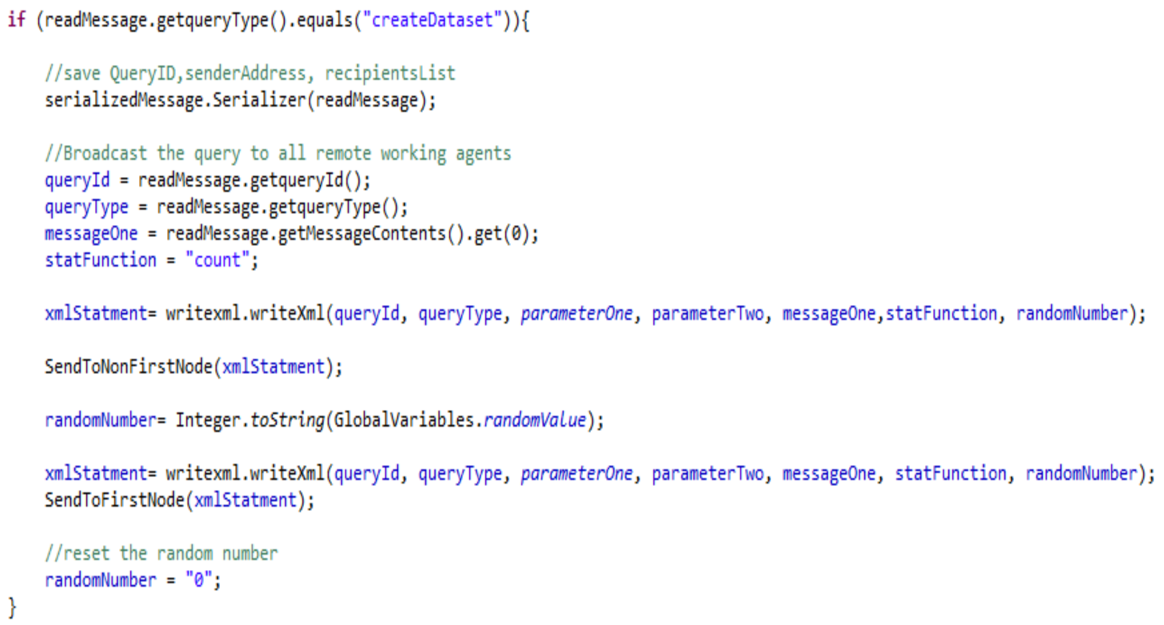


Figure 24. Data flow diagram of the Coordinating Agent.

The *Message Controller Module* is responsible for the major tasks of the *Coordinating Agent*. It manages the *Virtual Dataset* creation together with *Computing Agents*. It also implements the statistical computation graph described in chapter 5. *Statistical Computation Module* implements the local statistical computation required at the *Coordinating Agent*.

When a message is received at the *Communication Module,* the *Message Controller Module* uses the *Message Parser* (XML message reader) to parse the XML message and uses the parsed XML message to decide what to do next. If the message is a *Data Preparation* query coming from *Client Application*, the *Message Controller Module* broadcasts the query to the remote *Computing Agents* found at the health institutions identified by JIDs.



```
if (readMessage.getqueryType().equals("createDataset")){

    //save QueryID,senderAddress, recipientsList
    serializedMessage.Serializer(readMessage);

    //Broadcast the query to all remote working agents
    queryId = readMessage.getqueryId();
    queryType = readMessage.getqueryType();
    messageOne = readMessage.getMessageContents().get(0);
    statFunction = "count";

    xmlStatment= writexml.writeXml(queryId, queryType, parameterOne, parameterTwo, messageOne,statFunction, randomNumber);

    SendToNonFirstNode(xmlStatment);

    randomNumber= Integer.toString(GlobalVariables.randomValue);

    xmlStatment= writexml.writeXml(queryId, queryType, parameterOne, parameterTwo, messageOne, statFunction, randomNumber);
    SendToFirstNode(xmlStatment);

    //reset the random number
    randomNumber = "0";
}
```

Snippet 2. Message Controller Module example.

If the message is *Statistical Computation* query, the *Message Controller Module* uses the computation graph to jointly compute the Statistical function together with the *Computing Agents* and uses the *Statistical Computation Module (see Snippet 3)* for local computation. And it returns results to the *Client Application*.

6.4. Computing Agent

The *Computing Agent* is installed at every participating health institutions. It has three sub-components namely: *Communication, Message Controller, and Data Processing Modules*. The *Communication Module* sends and receives XMPP messages to/from Coordinating Agent, and other *Computing Agents*.

The *Message Controller* manages execution of queries in the *Computing Agent*. It uses the *Message Parser* to deserialize received XML messages into Java Object. The Java Object contains a query string, and other important information required for execution of a query.

The Data Processing Module, as illustrated in Figure 25, is where patient level data is processed. It executes AQL data preparation queries against the OpenEHR in phase I and SQL queries against the local data set in phase II. The two types of query executions are described in the following sub sections.

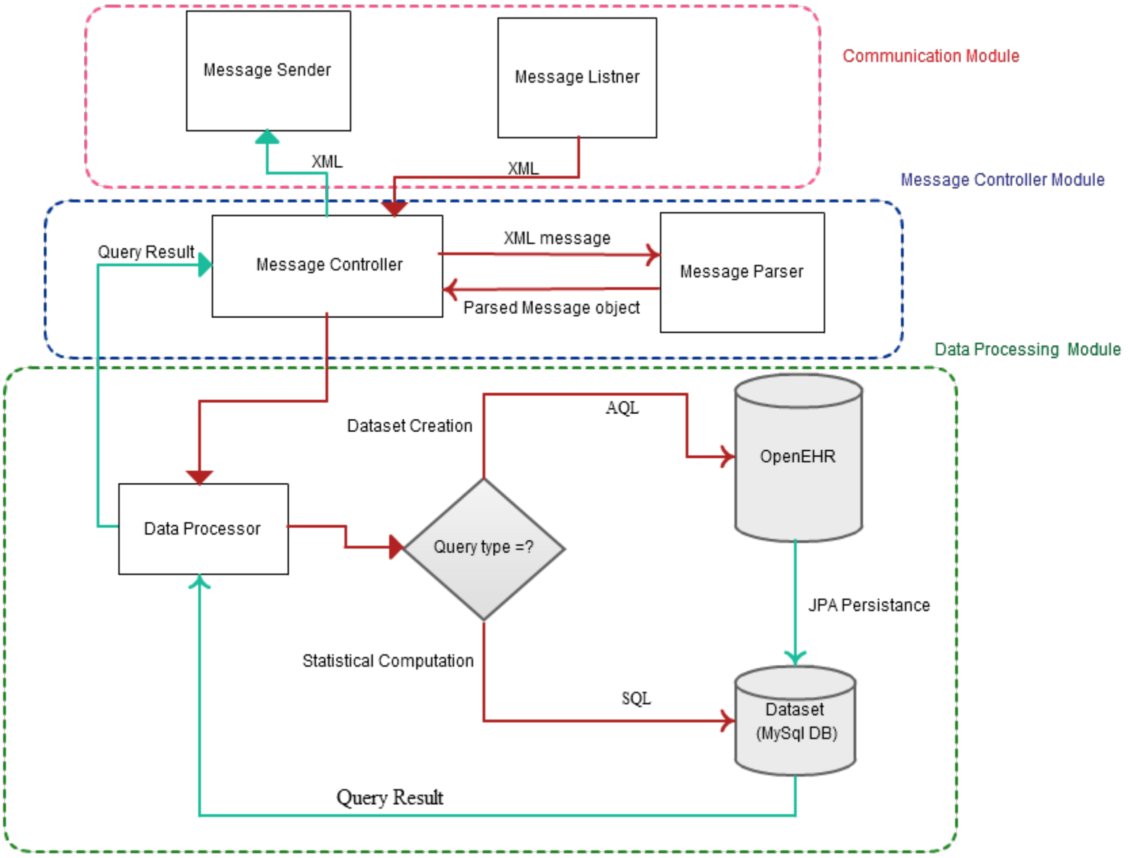


Figure 25. Data flow diagram of the Computing Agent.

6.4.1. Object/Relational Mapping

In phase I, the Data Processing Module used a Java API called "Thinkehr-Samples" that provides interfaces to execute AQL query against an OpenEHR repository. It returns the result as a Java Object. The Java Object is then persisted in a new MySQL relational database using Java Persistence API (JPA) object-relational mapping technique. JPA is a standard Java package, found in Java EE 5 and above, to manage object-relational mapping. Because it offers runtime persistence and useful development utilities, EclipseLink2.5 is used to implement JPA in this thesis.

The result of an AQL query execution at each health institution will have the same SQL schema, since the data are considered horizontally partitioned. The data sets (local MySQL database) at all participating health institutions cooperatively make data required for a research data set, called *Virtual Dataset*.

6.4.2. Virtual Dataset

Once the *Virtual Dataset* is created in phase I, the user can send as many statistical computation queries as needed. Since statistical computations are done on the *Virtual Dataset*, which is stored in relational databases across the health institutions, the Data processor module uses SQL query.

6.5. Summary

This chapter describes an implementation of the design described in chapter 5. We described the programming languages and technologies used for the implementation. Then we described the communication protocol we created for the message exchange among *Emnet's* components.

Then, the chapter explained the detail implementation of the three components of *Emnet:*

- Client Application
- Coordinating Agent
- Computing Agents

Chapter 7 Testing and Result

This chapter contains different sections that contribute to testing of the implemented prototype system. First, we describe specific scenarios that are collected from the users during requirements gathering meetings. The second and third sections describe the test data preparation process. The next section discusses the testing process and the final section presents the result obtained.

7.1. Scenarios

Scenario 1 – Scenario of Epidemiologist

As it became a major health problem worldwide, studies on obesity are increasing. In relation to this, a study (130) concluded that obesity and body temperature are inversely correlated in human. On the other hand, an article (131) recommended more research needs to be done in different areas and different circumstances. Hence, Dr. Maria wanted to test the correlation between human *body temperature* and *body mass index*.

Scenario 2 – Scenario of Clinical Researcher

As part of his research career, Dr. Robin wants to compute patients' Medical Complexity Score based on the primary care data all over Norway. Medical Complexity Score reveals the patients' complexity of healthcare needs and it is analyzed based on the following factors:

1. Number of unique diagnoses
2. Number of consultations with GP
3. Number of referrals from GPs
4. Number of referred unique departments/services
5. Number of discharge letters to unique institutions
6. Number of sick leaves
7. Number of laboratory tests
8. Number of prescriptions
9. Number of unique ATC codes

However, both users cannot use current data reuse solutions, because (1) the data sources do not share data, because of privacy; and (2) collecting large amount of data, such as the listed above, from distributed sources requires enormous amount of resources (time and money). Therefore, a privacy-preserving system that assists the researchers to perform the required computations is needed.

7.2. Test Data Preparation

In Norway, openEHR based EHRs are under development; so we could not find data that can be used for testing. Therefore, we prepared our own openEHR data to test *Emnet* with the scenarios described above. The following section has brief explanation of the test data preparation process.

7.2.1. Clinical Modeling

OpenEHR architecture uses two-level modeling that separates the reference information model and the clinical knowledge (concepts). Archetypes are used to define clinical concepts. In principle, archetypes are reusable components. However, since they are intended for clinical purpose, they are built to fit in the clinical concepts. Reusing archetypes (for some other purpose such as research) might require some modification to make them suitable for that special need. Hence, our scenarios require modification on some of the archetypes.

According to ***Scenario 1*** given above, the following two clinical concepts with their corresponding attributes are identified.

1. Body Temperature with attribute *Temperature*
2. Body Mass Index with attribute *Body mass Index*

According to ***Scenario 2*** given above, the following lists of clinical concepts with their corresponding attributes are identified.

1. Diagnosis (e.g. diabetes) with attributes *Problem/Diagnosis* and *Date of onset*.
2. Demography with attributes *Unique ID*, *Date of Birth* and *Sex*
3. Consultations with attributes *Reason for Consultation* and *Date of Consultation*
4. Referrals with attributes *Requester Id, Receiver Id, Referred Service* and *Date of Referral*
5. Sick leaves with attributes *Problem/Diagnosis* and *Start Date*
6. Laboratory test with attributes *Service Requested* and *Date of Request*
7. Prescription with attributes *Medicine* and *Date of Prescription*

7.2.2. Archetype Review

Basically, we used openEHR CKM in order to represent these clinical concepts in archetypes. The process of identifying the right archetype for each clinical concept was done in the following steps:

I. Type the keyword (the concept)
II. Go into the details of the more similar archetypes from the search results provided
III. Select the most similar archetype

IV. Check the availability of all attributes required for the specific scenario described above

V. List the missing attributes if there are any

VI. Edit the archetype by adding the missing attributes so that it can accommodate the specific need

VII. If openEHR CKM has no archetype for a concept, create a new archetype.

Hereafter, we explain the archetype preparation process of each clinical concept as follows:

Scenario 1.

1. Body Temperature
   Number of related archetypes found - 3
   Selected archetypes – *Body Temperature*
   Reason for selection – the *Body Temperature* archetype has the right attribute, *Temperature*, which is needed for the *Body Temperature* clinical concept in our scenario.
2. Body Mass Index
   Number of related archetypes found - 1
   Selected archetypes – *Body Mass Index*
   Reason for selection – the one and the needed archetype with the right attribute, *Body Mass Index*, which is needed for the *Body Mass Index* clinical concept in our scenario.

Scenario 2.

1. Diagnosis
   Number of related archetypes found - 32
   Selected archetypes – *Problem List* and *Problem/Diagnosis*
   Reason for selection – the *Problem/Diagnosis* archetype has many attributes including *Problem/Diagnosis* and *date of onset* which are the two attributes needed for the *Diagnosis* clinical concept in our scenario. In addition, there are cases when a patient is diagnosed for more than one disease. Therefore, we selected the *Problem List* archetype which has four slots (to plug in another archetype when needed). *Problems or Diagnoses* is one of the four slots with occurrence value of 0…* and it is used to plug in the *Problem/Diagnosis* archetype.
   Missing attribute - *none*
2. Demographic
   Number of related archetypes found - 40
   Selected archetypes – *Individual's personal demographics*
   Reason for selection – *Individual's personal demographics* archetype has many attributes including *identifier*, *Date of Birth* and *Sex* which are the three attributes

needed for the *Demographic* clinical concept in our scenario. Since all the necessary attributes in our clinical concept are found in this archetype, we selected this archetype.

Missing attribute - *none*

3. Consultation

   Number of related archetypes found - 19

   Selected archetypes – *Reason for Encounter*

   Reason for selection – *Reason for Encounter* archetype has *Reason for Contact* attribute which is needed for the clinical concept Consultation in our scenario.

   Missing attribute – *Date of Consultation*

4. Referrals

   Number of related archetypes found - 6

   Selected archetypes – *Referral*

   Reason for selection – *Referral* archetype has many attributes including *Requester Identifier, Receiver Identifier and Referred Service* which are three of the four attributes needed for the clinical concept *Referral* in our scenario.

   Missing attribute – *Date of Referral*

5. Sick leaves

   Number of related archetypes found - 0

   Selected archetypes – *none*

   Archetype creation Process – Using the openEHR Archetype Editor we created a new archetype as it fits our scenario. Thus, the archetype name is called *Sick leave* and it has the following five attributes: *Id* (with *Text* data type)*, Problem/Diagnosis* (with *Text* data type)*, Start Date* (with *Date* data type)*, End Date* (with *Date* data type) and *Comment* (with *Text* data type).

6. Laboratory test

   Number of related archetypes found - 40

   Selected archetypes – *Laboratory Test Request*

   Reason for selection – *Laboratory Test Request* archetype has *service requested* attribute which is needed for the *Laboratory Test* clinical concept in our scenario.

   Missing attribute – *Date of Request*

7. Prescription

   Number of related archetypes found - 4

   Selected archetypes – *Medication Action*

   Reason for selection – the *Medication Action* archetype has many attributes including *Medicine* which is an attribute needed for the *Prescription* clinical concept in our scenario.

   Missing attribute – *Date of Prescription*

7.2.3. Archetype Editing

As we can see from the section above, some of the archetypes found in the openEHR CKM have missing fields for the specified clinical concept in our particular scenario. Consequently, those archetypes require editing. Archetype Editor (132) is used to edit the existing archetypes and to create new archetypes. OpenEHR Archetype Editor is a software provided by Ocean Informatics[11] to develop and edit archetypes both for clinical and research purpose.

To edit the archetypes, first we downloaded the identified archetypes (in section 5.2.1.1 above) from openEHR CKM. Then, we used the Archetype Editor tool to edit the archetypes with missing attributes. The editing was basically creating extra fields in the archetype that represent the missing attributes. In doing so, the edited archetypes fit in the given scenario.

The Archetype Editor is also used to create a new archetype - *sick leave*. Since this archetype is not found in the openEHR CKM yet, we created a new archetype with the attributes needed for the scenario.

7.2.4. Template Design

After collecting all required archetypes, we designed two templates (for ***Scenario1*** and ***Scenario2***) namely *BMI_BT Correlation* and *Patient Summary* as shown in Figure 26 *and* Figure 27 respectively. The Ocean Informatics' Template Designer [112] is used to design the templates by drag-and-drop the archetypes into the template stem.

The *BMI_BT Correlation* template represents the set of data in use case ***Scenario 1***.

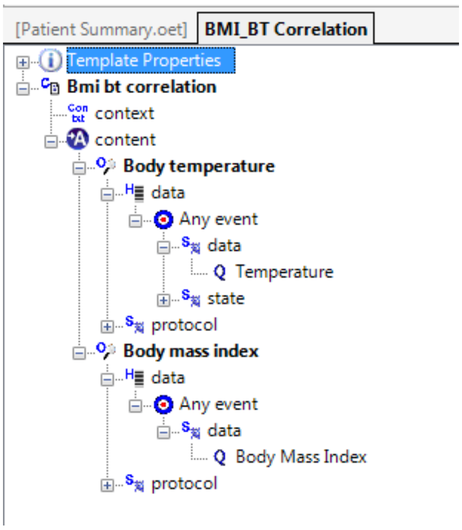
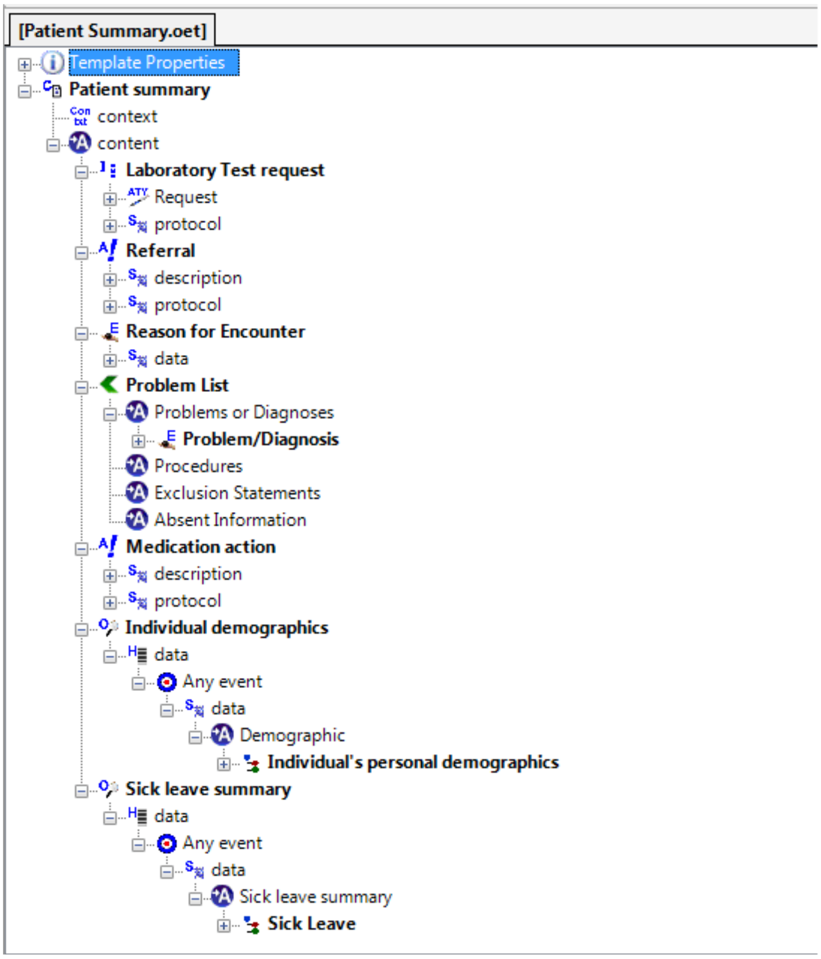


Figure 26. BMI_BT Correlation template containing the selected archetypes in Scenario 1.

[11] https://oceaninformatics.com/solutions/knowledge_management

The *Patient Summary* template represents the set of data in use case ***Scenario 2***.



[Patient Summary.oet]

- Template Properties
- Patient summary
  - context
  - content
    - Laboratory Test request
      - Request
      - protocol
    - Referral
      - description
      - protocol
    - Reason for Encounter
      - data
    - Problem List
      - Problems or Diagnoses
        - Problem/Diagnosis
      - Procedures
      - Exclusion Statements
      - Absent Information
    - Medication action
      - description
      - protocol
    - Individual demographics
      - data
        - Any event
          - data
            - Demographic
              - Individual's personal demographics
    - Sick leave summary
      - data
        - Any event
          - data
            - Sick leave summary
              - Sick Leave

Figure 27. Patient Summary template containing the selected archetypes in Scenario 2.

7.3. Test Data Creation

Normally, *Emnet* is developed with the assumption of using openEHR based EHRs as data sources. However, in reality this kind of data is not widely available as it fits our testing plan. Therefore, generating openEHR data used in testing *Emnet* was one of the tasks done during the testing phase.

The test data creation process highly depends on a software tool (library) provided by Marand. The tool uses a template made of the required archetypes and some other specifications to produce set of data in openEHR format. Therefore, we created the test data only for "***Scenario 1***". We followed the steps described above (clinical modeling, archetype review, archetype editing and template designing) to prepare the template to be used by the Marand's tool and finally created the archetype-based data.

7.4. The Testing Procedure

The testing procedure was performed on virtual distributed environment established for the testing purpose. Following is the explanation of the virtual testing environment setup.

Lenovo X230 ThinkPad (windows 7 OS) computer is used to setup the virtual testing environment. The technologies used to establish the virtual distributed testing environment include:

- VirtualBox 4.3.20[12]
- Vagrant 1.7.2[13]
- Ubuntu14.04[14] operating system.

VirtualBox is a general-purpose virtualization tool provided by Oracle. It enables to create virtual machines on top of the existing operating system.

First, the Oracle VirtualBox is installed on the host machine (the computer) then a Vagrant script was used to create three instances of GP office computer (named it GP1, GP2 and GP3) and one Coordinating server. The Vagrant script also installed Ubuntu OS and other (such as MySQL, Java) necessary software on the guest machines. The Vagrant script also establishes all the required networking configurations.

The machines involved in the testing and the installed tools are described below.

Host Machine

- Openfire server
- Apache web server to host the web client application
- Client Application

Virtual GP office computers (GP1, GP2, and GP3)

- OpenEHR instance (Think!EHR repository)
- Computing Agent component

Coordinating Server

- Coordinating Agent component

Hereafter, we typed the web address of the *Client Application* in Firefox on the host machine and the login page is displayed. After typing the correct Jabber ID (JID) and password, it

[12] https://www.virtualbox.org/

[13] https://www.vagrantup.com/

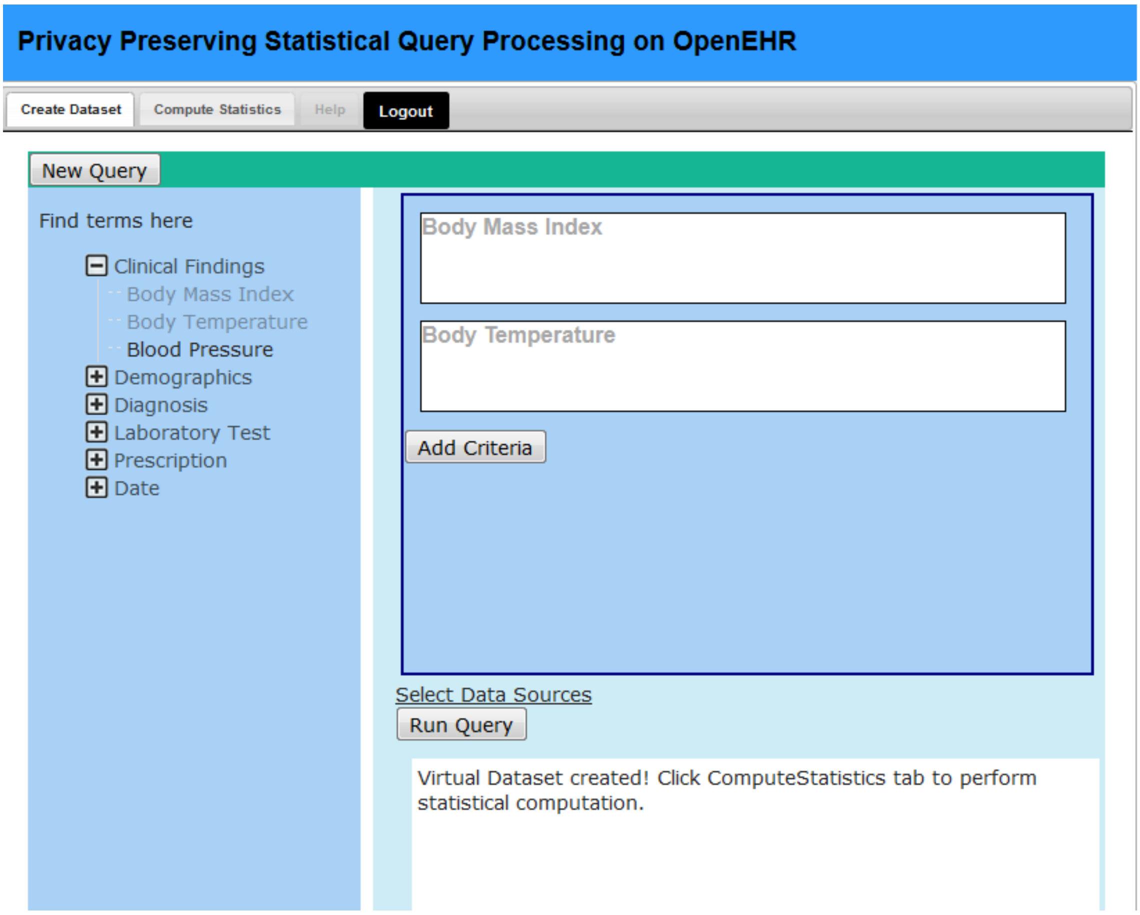
[14] http://www.ubuntu.com/index_roadshow

authenticated on the openfire server. Consequently, the *Data Preparation* page of the *Client Application* is displayed.

7.5. Result

As it fits ***Scenario 1*** (see section 7.1.), we used two medical concepts to specify the research criteria. The medical concepts used are Body Mass Index and Body Temperature and the step-by-step criteria specification procedure is described using Figure 28 as follows.

1. Click the "Add Criteria" button – a white box and a dialog box will appear (see Appendix C1)
2. Select inclusion /exclusion option for the criteria from the dialog box
3. Drag "Body Mass Index" and drop it on the white box - a dialog box will appear (see Appendix C2)
4. Set a value/range of value for Body Mass Index on the input box of the dialog box
5. Repeat step 1 to 4 for "Body temperature"
6. Click the "Run Query" button



Virtual Dataset created! Click ComputeStatistics tab to perform statistical computation.

Figure 28. Client Application during research criteria specification phase.

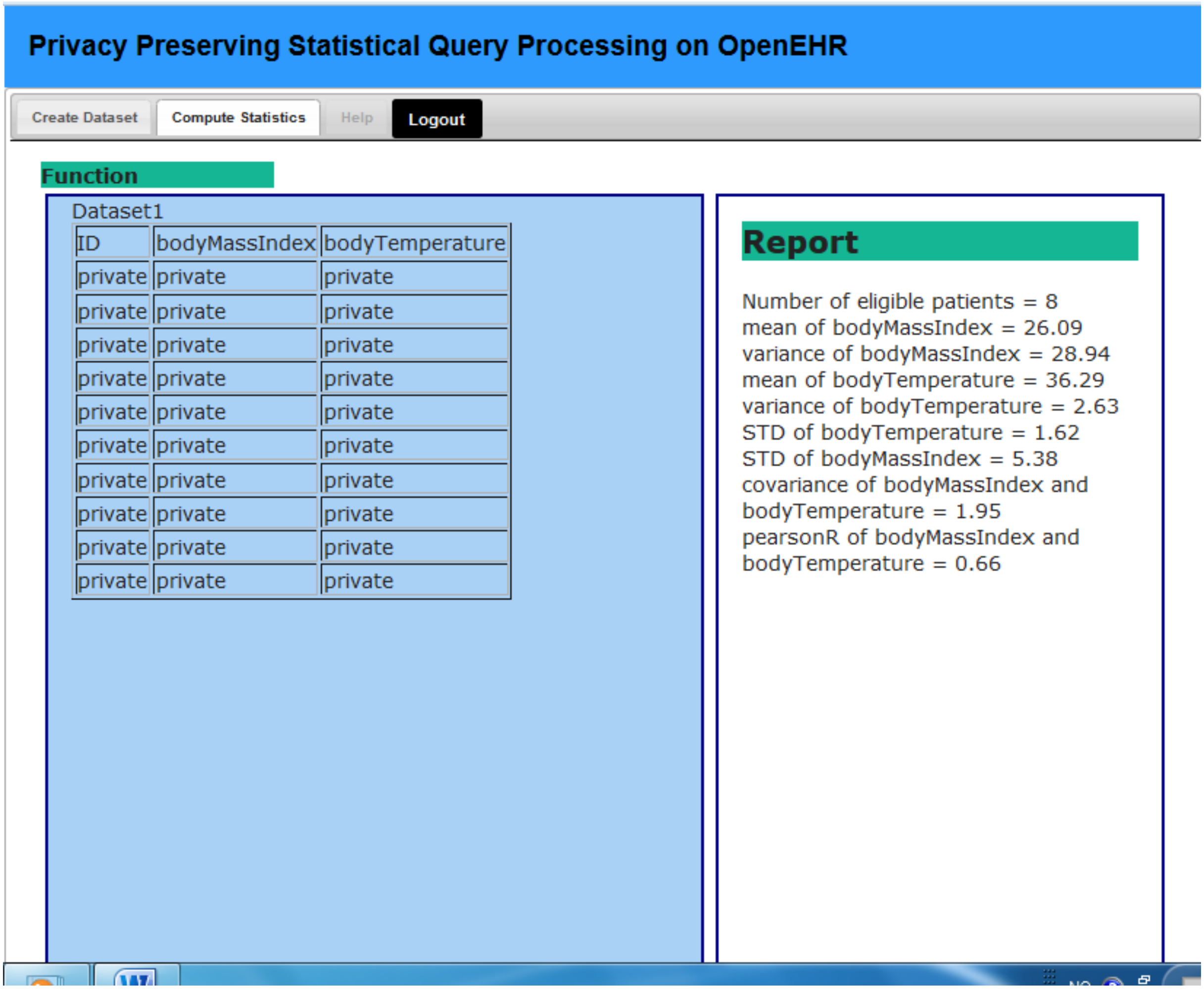
7. A notification message is displayed on the white area below the "Run Query" button

8. As shown in Figure 28, the success message is displayed when the *Virtual Dataset* is created.

Hereafter, by clicking the "Compute statistics" tab, a page indicated in Figure 29 is displayed. The page has the tabular description of the *Virtual Dataset* and also the number of eligible patient for the specified research criteria (*Count*) on the report section, which is found on the right side of the page.

A list of statistical functions is available in the "Function" menu. The following steps are used to perform statistical computations.

1. Click on the "Function" menu
2. Select the "Mean" sub menu – a dialog box will appear (see Appendix C3)
3. Select a variable (of which the ***mean*** is to be computed) from the dialog box
4. Click "Compute" button
5. Repeat the same steps to compute the value of all the required statistical functions



Dataset1

ID	bodyMassIndex	bodyTemperature
private	private	private
private	private	private
private	private	private
private	private	private
private	private	private
private	private	private
private	private	private
private	private	private
private	private	private
private	private	private

Report

Number of eligible patients = 8
mean of bodyMassIndex = 26.09
variance of bodyMassIndex = 28.94
mean of bodyTemperature = 36.29
variance of bodyTemperature = 2.63
STD of bodyTemperature = 1.62
STD of bodyMassIndex = 5.38
covariance of bodyMassIndex and bodyTemperature = 1.95
pearsonR of bodyMassIndex and bodyTemperature = 0.66

Figure 29. Client Application displaying metadata information and statistical computation report.

The result of the computed values displayed under the “Report” section. All the results of the computations are listed sequentially as shown in Figure 29. During this testing, five types of statistical functions are tested using ***Scenario1***. Therefore, the report is made of *Count* (the number of eligible patients) and the computation result of the five statistical functions. These are: *Count, Mean, Variance, STD (Standard Deviation), Covariance,* and *Pearson’s r.*

7.6. Summary

The chapter discussed the series of tasks done to test the prototype system. The first section describes the use case scenarios provided by the potential users of *Emnet*. The next section presented how the data used for testing was prepared. It explained the archetype-based data preparation for the given scenarios, which involves the following steps.

- Clinical Modeling
- Archetype Review
- Archetype Editing
- Template Design

The succeeding section told how the data used for testing was created out of the prepared templates. It also described the role of Marand’s openEHR data creation library in the data creation process. Further down, the chapter briefly described the testing procedure and how the distributed virtual environment was prepared for the testing. Finally, the chapter concluded by describing the results of the testing procedure. It presented the results obtained at every step of the testing process.

Chapter 8 Discussion

This chapter summarizes the thesis and discusses the main techniques and their interpretations. It provides a short summary of the thesis along with justifications of the main approaches used. Then, the thesis findings are discussed in comparison with related works. Finally, limitations of the thesis are discussed.

8.1. The Proposed System

The thesis designed a system that enables researchers (users) perform statistical computations on data distributed across health institutions, while preserving the privacy of patients' and health institutions. *Emnet* is designed as per the requirements specified in chapter 4 and contains three components: *Client Application*, *Computing Agent*, and *Coordinating Agent*. The *Client Application* is a web application where users interact with *Emnet*. The *Computing Agent* is installed at each health institution. It computes on local data and performs secure summation protocol for joint privacy-preserving computation with other *Computing Agents*. The *Coordinating Agent* is located between a *Client Application* and *Computing Agents*. It coordinates execution of user's queries on data located at multiple health institutions in a privacy-preserving manner. These components are loosely coupled through XMPP messages. As a result, one component can be modified or changed without affecting the other components.

Emnet has research *Data Preparation* and *Statistical Computation* phases. The two phases are described in the following sub-sections.

8.1.1. Phase I – Data Preparation

According to the security requirement, the *Client Application* authenticates users with JID and password. Only then a user can access the interface to specify criteria that select research data set (*Virtual Dataset*). User's criteria are compiled into AQL query before sent to the *Coordinating Agent*. The choice of AQL to express user's criteria as computable criteria is based on the assumption that all health institutions use openEHR-based EHRs. We will discuss the basis for this assumption later in this section. Since *Emnet* is assumed to compute on horizontally partitioned data, the *Coordinating Agent* uses the following simple steps, (1) it broadcasts the AQL query to *Computing Agents* at *Participating Health Institution*, and (2) each *Computing Agent* executes the AQL query against its EHR and locally stores the query result in a MySQL relational database, which stays for approved research duration.

One might ask, if the query results are not going to move outside the health institutions, why then store in a separate relational database? Because of the following reasons we chose to use a separate database: (1) to prevent any modification on the EHR that might occur during statistical computation; (2) to facilitate EHR data cleaning and transformation to make the data suitable for the research and (3) to avoid frequent access to the original EHR for every statistical computation requests.

Since users will not see the *Virtual Dataset*, *Emnet* displays metadata and descriptive statistics (e.g. count of eligible patients) to help the user visualize it. It also helps to know what kinds of data variables are available to perform further computations in the next phase.

8.1.2. Phase II – Statistical Computation

Emnet enables privacy-preserving Statistical Computations on a *Virtual Dataset*. The conventional health data reuse requires either patient consent or data de-identification. These methods are useful and widely used, yet there are also counter arguments that are discussed in chapter 2. In contrast, *Emnet* enables data reuse, without data de-identification or consent, while providing better privacy guarantee. The technique used to achieve this requirement is described below.

A user sends statistical computation query via the *Client Application* to the *Coordinating Agent*. The *Coordinating Agent* coordinates privacy-preserving execution of the query on the *Virtual Dataset* and returns the result to the user using the following techniques. Primarily, it uses a computation graph that describes decomposition of statistical functions into sub-computations of summation forms. It coordinates execution of each sub-computation in two rounds, (1) the *Computing Agent* at each health institution computes the sub-computation on its local data set; and (2) all *Computing Agents* across health institutions jointly sum their local results using secure summation protocol. The current implementation used SINE secure summation protocol (103), mainly because it balances between the privacy guarantee and computation complexity among the other reviewed protocols. However, *Emnet*'s design allows to easily replace it with a protocol that have similar or better privacy guarantee.

The local computation hides patients' information by aggregating multiple patients' data together and releasing it does not violate patients' privacy. Yet, it can be considered private information for the institution. Then, the secure summation aggregates an institution's private information within multiple institutions information. Therefore, the technique enables to achieve the privacy requirements of both patients and health institutions.

Apart from ensuring privacy, the nature of *Emnet* offers the health institutions with the right to see the query to be executed against the data (EHR) their organization is providing. Hence, the patients and health institutions may feel more confident to give data for research which in turn increases research data availability.

SMC is considered computationally expensive. However, our approach is efficient and practical, because sub-computations are executed in parallel at each health institution on individual level data. Therefore, it leverages the distribution of the data. In addition, secure summation protocol is only used to sum sub-computation results of the health institutions, and it is much efficient than other SMC protocols (96).

8.1.3. How will *Emnet* be used?

The main potential users of *Emnet*, identified in chapter 4, are epidemiologists, clinicians (i.e. general practitioners (GPs)), and clinical researchers. However, in its current implementation, it mainly focused on epidemiologists and clinical researcher. Epidemiologists and clinical researchers can use *Emnet* for one time query to find eligible patients or research. In the latter case, they create research data sets that last for the project's time period and execute statistical computation on the data sets. However, the basic design of *Emnet* also supports the requirements of clinicians.

GPs would like to benchmark their clinical practice performance with respect to the average performance of GPs in their area. Privacy-preserving benchmarking needs to compare statistical computation on all health institutions data set (*Virtual Dataset*) and a GP's local data set. Computations on a particular GP's local data set can only be performed inside the GP office's local area network (LAN). However, since other potential users cannot see computation result of a single institution, currently, the *Client Application* only communicates with the *Coordinating Agent* and gets computation results on a *Virtual Dataset.* Therefore, to support privacy-preserving computation on a single GP office data set, the *Client Application* needs to directly communicate with the *Computing Agent* at the office of the GP requesting the benchmark; and the *Computing Agent* computes on the local data set and return the result.

8.2. Comparison to Related Works

The number of publications on privacy-preserving health data reuse indicates how much focus has been given to the topic. However, only a handful of distributed EHRs query tools and distributed research networks that are related to *Emnet* have been identified and discussed in the literature review section of chapter 2. The reviewed systems are SAFTINET (95), EHR4CR (9), SHRINE (14), PopMedNet (17), and SCANNER (18).

These are big projects and have very wider focus than this thesis. Therefore, the discussion and comparison is limited to techniques used for different steps of privacy-preserving statistical computation, such as common data model, and implemented statistical functions.

The current strategic plan of Norwegian Health authorities is encouraging EHR vendors to adopt openEHR (23). Based on this direction, we assumed that health institutions use openEHR based EHRs and it is used as common data model for *Emnet*. In contrast, all the other systems use a different data model than the source EHR. As a result, they transform data at each data source to their common data model before reuse. For instance, SHRINE has a four-step data transformation process in order to achieve semantic interoperability when using the query tool. When the health institutions use openEHR based EHRs, no extract, transform and load (ETL) process is needed.

Except *Emnet* and SCANNER, to our knowledge, the other systems only implemented *count* of eligible patients. The current implementation of SCANNER contains statistical functions such as *mean, standard deviation, and binary logistic regression*, while *Emnet* implemented *mean, variance, Standard Deviation, Covariance* and *Pearson's r.* The privacy-preserving statistical computation technique described in the thesis is generic and can easily be extended for more statistical functions (see Appendix B for list of statistical functions that can easily be implemented).

Except *Emnet* and SCANNER, the other systems seem to be concerned about only the privacy of the patients, but not the health institutions'. As a result, they release individual institution level computation results. But, in reality privacy is not only the concern of patients but also of health institutions and clinicians (47). The current implementation of SCANNER also does not protect the privacy of health institutions. However, they have proposed a solution to include privacy protection of health institutions. The techniques presented in this thesis protect the privacy of both patients and health institutions.

8.3. Limitations

Because of the sensitive nature of patient data, EHR data reuse is limited. Hence, it needs complex techniques and extensive amount of time to implement a solution. In this thesis, we developed privacy-preserving technique to perform different statistical computations on distributed EHRs. However, it is wise to consider the limitations before implementing *Emnet* in the real life environment. The major limitations of *Emnet* are discussed below.

Improved Security

Security is one of the non-functional requirements, and *Emnet* implemented user authentication mechanism to meet this requirement. Yet, it requires other security features. Snow system has strong security mechanisms, such as authentication using certificates and end-to-end message encryption between two communicating nodes. Since *Emnet* is part of the Snow system, the plan is to integrate the two systems. Therefore, *Emnet* will reuse the Snow's security facilities; and the current implementation did not give much focus on the security aspects.

Scalability

Scalability is one of the main aspects to consider in distributed systems. Because, the performance of a system is affected as the number of nodes connected to the system grows. In this thesis, the testing was performed with three data sources and it was therefore difficult to test the scalability issues. Hence, some secure-scalability features should be considered to implement *Emnet* in production environment. For example, A paper (106) suggested secure-parallel computation where the whole computation is divided and handed to different groups of neighbor nodes to compute their values secretly and aggregate results in parallel.

Fault-tolerance

Emnet works on distributed systems where the failure of one component affects the overall computation process. For example, if a single component at one of the data sources failed to perform the expected task, *Emnet* should be able to continue working with the rest of the components. However, the fault-tolerance mechanism should be intelligent enough to avoid statistical errors that might occur in cases of missing values due to system failure. Faults can happen due to many reasons including hardware failure, software failure, and network failure. Even, a health institution might decide to stop sharing data in the middle of a computation. We were not able to implement strong fault tolerance mechanisms in this thesis.

Data Related Issues

In this thesis, we had the assumption that current EHRs data are suitable to be used directly for research. However, in practice, rigorous work has to be done to correct and prepare the data to make it suitable for research. Data quality, duplicate data and data partitioning are among the identified data related issues that might limit data reuse.

Data quality is one of the main challenges of EHR data reuse. Because, EHR data (data collected for patient treatment) may not fit the research need. A study (27) identified completeness, correctness, concordance, plausibility, and currency as the common measurements of data quality.

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. Therefore, duplicate records could occur across institutions where patients received treatment.

The focus of this thesis is only horizontally partitioned data. The techniques developed in the thesis needed modification to enable computation on vertically partitioned data.

8.4. Summary

This chapter summarized the thesis and discussed the main techniques developed in the thesis. The first section discussed the developed solution with the explanations of the approaches used. The next section presented the techniques developed in the thesis in comparison with related works identified in the literature review section of chapter 2. The last section described the limitations of the thesis and the main issues that should be addressed in order to make the proposed solution fully functional in real production environment.

Chapter 9 Conclusion and Future Work

9.1. Conclusion

In this thesis we developed a framework and implemented a prototype of privacy-preserving statistical computation system on distributed health data. *Emnet* was developed to satisfy requirements specified in chapter 4. In general, the thesis demonstrated a practical solution for data reuse without a need to collect the data in a centralized repository, while ensuring strong privacy protection. The *Virtual Dataset* creation, and statistical computation technique used in the thesis provides direct control to the data owners what/who compute on their data.

In addition to re-identification risks vs. data utility arguments, data de-identification methods are limited to protect the privacy of health institutions. Addressing the privacy concerns of health institutions can increase health institutions' and individuals' willingness to share more data for research.

The work presented in this thesis is also relevant and can be used in other domains, such as finance where a need for joint computation on data of multiple data sources contradicts with privacy requirements.

In the following paragraphs, we briefly discussed the solutions developed in the thesis to answer each research question.

> ***Question 1.*** *How can the research inclusion and exclusion criteria be specified?*

A user-friendly web application is designed to help users easily specify inclusion/exclusion criteria and required attributes by drag and drop of list of medical concepts. Then, archetypes from the Norwegian archetype repository[10] are used to build an AQL query from the selected medical concepts.

> ***Question 2.*** *How can a research data (a Virtual Dataset) be created based on research criteria without moving the data outside the health institutions?*

Since *Emnet* assumed horizontally partitioned data*,* an AQL query is broadcasted to all participating health institutions. The institutions execute the query against their openEHR-based EHRs and locally store the resulting data set in a separate relational database. All the data sets, while remaining at the health institution they belong, jointly form a complete research data set (*Virtual Dataset)*.

> ***Question 3.*** *How can statistical computations be performed on the distributed data sets in a privacy-preserving manner?*

We developed a privacy-preserving statistical computing technique on top of the *Virtual Dataset* by leveraging (1) the decomposability of several statistical functions into sub-computations of summation forms; and (2) availability of efficient secure summation protocols.

The decomposition of the statistical functions into sub-computations and dependencies between them is described as a computation graph. Then, the *Coordinating Agent* implements it as computable objects. For each statistical function, the *Coordinating Agent* uses the computation graph to find the required sub-computations and their order of execution. A sub-computation might use preceding sub-computations' results.

A sub-computation is executed in two rounds, (1) each health institution locally executes the sub-computation on its local data set, and stores the result locally (sensitive intermediate value); (2) the health institutions jointly sum the intermediate values using secure summation protocol (the current implementation uses a protocol called SINE).

The practicality of the framework is demonstrated by the statistical functions implemented in *Emnet*, such as *mean, variance, Standard Deviation, Covariance* and *Pearson's r*.

9.2. Thesis Contribution

First privacy-preserving Statistical query tool for distributed EHRs in Norway

The literature review search covered well-known international databases in order to understand the state-of-the-art and identify related works. Most of the existing tools are designed to work in a centralized approach (where they centrally collect identifying or de-identified individual patient record). On the other hand, few tools are designed to work in a distributed manner to support a better privacy feature, but these tools focus on patient's privacy. However, as described in the literature review, privacy is not only the concern of individual patients' but also of health institutions'.

To the best of our knowledge, no distributed privacy-preserving statistical query tool is developed for EHRs in Norway. Therefore, we claim that this is the first attempt to develop a tool used to perform statistical computations on distributed EHRs while the preserving privacy of both individual patients' and individual health institutions'.

A Design that supports Interoperability

As discussed in the first chapter, interoperability is also among the major challenges of data reuse. However, a national openEHR archetype repository[10] is being developed to enable interoperability among EHRs in Norway. Considering this, our solution is designed using openEHR specifications as a common data model. Therefore, our implementation demonstrates significance of openEHR based EHRs for data reuse in addition to clinical benefits. Consequently, it could play a role to push forward currently started steps towards openEHR based EHRs in Norway.

Scientific Publication

A paper entitled "***Privacy-preserving Statistical Query and Processing on Distributed openEHR data***" (*Meskerem Asfaw Hailemichael, Luis Marco Ruiz and Johan Gustav Bellika*) is accepted in Medical Informatics Europe Conference (MIE 2015)[2]. The paper mainly contains the high level description of the solution together with the architecture developed in the thesis (see Appendix D).

Reusable Objects and Components

The literature reviews of privacy-preserving health data techniques and secure summation protocols provide good overview of the state-of-the-art on these topics. Hence, they can be basis for research in the field of data reuse. Moreover, one of the design considerations of *Emnet* is reusability. The objects and components used in *Emnet* are made to be reusable and easily extensible. Thus, the design and architecture can be reused for future development.

9.3. Future Work

The following are the major ideas that we consider are important to deal with in the future in order to produce a better version of *Emnet*. Some of the points are described as the limitations of the thesis.

More Statistical Functions

Even if we implemented selection of eligible patients and five functionalities of statistical computation in the current system, additional statistical functions are required to provide a full stack of statistical analysis capability for researchers. Hence, implementation of other essential statistical functions should be one of the future works to make *Emnet* more useful. The technique developed and implemented in this thesis enable easy integration of more statistical functions, as long as they can be decomposed into sub-computations of summation forms (see Appendix B).

Scalability and Fault-tolerance Techniques

Both *scalability* and *fault-tolerance* are the major concerns in distributed systems. Likewise, *Emnet* needs some more efforts in order to support these features. *Scalability* can be achieved by implementing parallel processing techniques to preserve the performance of *Emnet* when the number of data sources increases. Techniques such as Failure Resilient Processes (masks failures and guaranties progress despite the failure) can be used to achieve *Fault-tolerance.* However, great care should be taken because inappropriate inclusion or exclusion of input values may bring a huge difference in the final result.

Extensive Requirements Analysis

Basically, the sources of the requirements specification in this study are the literature review and the interviews with users. However, it is believed that the more users are involved in the requirements analysis the better the system will be. Moreover, *Emnet* is designed to support possible extensions provided that privacy is not violated. Therefore, one future work direction can possibly be further analysis of additional users' requirements.

Data Related Issues

All the data related issues discussed in *section 8.3.* (i.e. data quality, data partitioning and data duplication) are critical for data reuse. Therefore, handling these concerns is essential as future work.

Different data cleaning mechanisms are introduced to facilitate data reuse. Different privacy-preserving de-duplication techniques are reviewed in (134).

Testing Emnet in a Real Environment

Emnet is tested in a virtual environment where four distributed nodes (one coordinator and three health institutions) are simulated on a single computer. Everything seems pretty good as all the nodes reside on the same PC. However, issues such as network quality, working capacity of the computers and database sizes of data sources might affect the performance of the whole system in reality. Therefore, in the future, performance measurement needs to be done in a real environment.

Integration with the Snow system

Considering future integration, *Emnet* has Snow system inspired architecture. Thus, *Emnet* can reuse some of the functionalities in the Snow system. The functionalities include:

- Selecting data sources – the address of the participating health institutions
- Broadcasting queries – the message exchange facilities
- Enhanced security - authentication and end-to-end message encryption features

Consequently, *Emnet* will benefit from the above and other features of the Snow system following future integration.

Better Protocol for Privacy

The protocol implemented in the prototype works based on the assumption that the *Coordinating Agent* is semi-trusted, it will not collude with Computing Agent. However, a more sophisticated protocol may potentially be implemented to achieve stronger privacy without depending on the trustworthiness of any components of *Emnet* (135).

References

1. Bloomrosen M, Detmer D. Advancing the Framework: Use of Health Data--A Report of a Working Conference of the American Medical Informatics Association. J Am Med Inform Assoc. 2008;15(6):715–22.
2. Dentler K, ten Teije A, de Keizer N, Cornet R. Barriers to the reuse of routinely recorded clinical data: a field report. Stud Health Technol Inform. 2013;192:313–7.
3. Selby JV, Krumholz HM, Kuntz RE, Collins FS. Network News: Powering Clinical Research. Sci Transl Med. 2013 Apr 24;5(182):182fs13–182fs13.
4. Friedman CP, Wong AK, Blumenthal D. Achieving a Nationwide Learning Health System. Sci Transl Med. 2010 Nov 10;2(57):57cm29–57cm29.
5. Randhawa GS, Slutsky JR. Building Sustainable Multi-functional Prospective Electronic Clinical Data Systems: Medical Care. 2012 Jul;50:S3–6.
6. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care. 2010 Jun;48(6 Suppl):S45–51.
7. Berlin JA, Stang PE. CLINICAL DATA SETS THAT NEED TO BE MINED. Learning What Works: Infrastructure Required for Comparative Effectiveness Research: Workshop Summary [Internet]. National Academies Press (US); 2011. Available from: http://www.ncbi.nlm.nih.gov/books/NBK64781/
8. Weiner MG, Embi PJ. Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning? Ann Intern Med. 2009 Sep 1;151(5):359–60.
9. Ouagne D, Hussain S, Sadou E, Jaulent M-C, Daniel C. The Electronic Healthcare Record for Clinical Research (EHR4CR) information model and terminology. Stud Health Technol Inform. 2012;180:534–8.
10. Rothstein MA. Is Deidentification Sufficient to Protect Health Privacy in Research? Am J Bioeth. 2010 Sep;10(9):3–11.
11. Malin BA, Emam KE, O'Keefe CM. Biomedical data privacy: problems, perspectives, and recent advances. J Am Med Inform Assoc. 2013 Jan 1;20(1):2–6.
12. HHS awards $1M contract for effectiveness database | Government Health IT [Internet]. [cited 2014 Feb 10]. Available from: http://www.govhealthit.com/news/hhs-awards-1m-contract-effectiveness-database
13. Maro JC, Platt R, Holmes JH, Strom BL, Hennessy S, Lazarus R, et al. Design of a National Distributed Health Data Network. Ann Intern Med. 2009 Sep 1;151(5):341–4.
14. Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. J Am Med Inform Assoc. 2009 Sep 1;16(5):624–30.
15. Lazarus R, Yih K, Platt R. Distributed data processing for public health surveillance. BMC Public Health. 2006 Sep 19;6(1):235.
16. Research Patient Data Registry (RPDR) | Research Information Services & Computing [Internet]. [cited 2014 Feb 26]. Available from: http://rc.partners.org/rpdr
17. PopMedNet | Distributed Research Network Technologies for Population Medicine [Internet]. [cited 2014 Sep 8]. Available from: http://www.popmednet.org/
18. SCANNER | Scalable National Network for Effectiveness Research [Internet]. Available from: http://scanner.ucsd.edu/
19. Hailemichael MA. Privacy preserving statistical query on distributed health data. UiT- The Arctic University of Tromso; 2014 Jun.
20. Richesson RL, Krischer J. Data Standards in Clinical Research: Gaps, Overlaps, Challenges and Future Directions. J Am Med Inform Assoc. 2007 Nov 1;14(6):687–96.
21. Lau LM, Shakib S. Towards Data Interoperability: Practical Issues in Terminology Implementation and Mapping. HIC 2005 and HINZ 2005 [Internet]. 2005 [cited 2015 May 11]. Available from: http://search.informit.com.au/documentSummary;dn=994111375505125;res=IELHEA
22. HIMMS | Transforming health through IT [Internet]. 2013. Available from: http://www.himss.org/library/interoperability-standards/what-is
23. Ellingsen G, Christensen B, Silsand L. Developing Large-scale Electronic Patient Records Conforming to the openEHR Architecture. Procedia Technology. 2014;16:1281–6.
24. DIPS: Customers [Internet]. Available from: http://www.dips.no/eng/about-us/customers?lang=eng

25. DIPS: DIPS Arena Client [Internet]. DIPS Arena Client. Available from: http://www.dips.com/eng/our-solutions/dips-arena-client?lang=eng
26. What is openEHR? [Internet]. [cited 2014 Mar 6]. Available from: http://www.openehr.org/what_is_openehr
27. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan 1;20(1):144–51.
28. 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 [Internet]. Nass SJ, Levit LA, Gostin LO, editors. Washington (DC): National Academies Press (US); 2009 [cited 2015 May 5]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK9578/
29. Berman JJ. Confidentiality issues for medical data miners. Artif Intell Med. 2002 Oct;26(1-2):25–36.
30. Toll K. Privacy and Freedom. By Alan F. Westin. New York: Atheneum Press, 1967. $10.00. Social Work. 1968 Oct 1;13(4):114–5.
31. Resources - Data Privacy [Internet]. INTERNATIONAL SOCIETY FOR PHARMACOEPIDEMIOLOGY (ISPE). [cited 2014 Sep 18]. Available from: https://www.pharmacoepi.org/resources/privacy.cfm
32. Hayrinen K, Saranto K, Nykanen P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. Int J Med Inform. 2008 May;77(5):291–304.
33. EuroREACH: Access to National Health Data Systems [Internet]. Available from: http://www.euroreach.net/activities/workpackages/wp3
34. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. J Intern Med. 2013 Dec 1;274(6):547–60.
35. ACT 2008-06-20 no. 44: Act on medical and health research (the Health Research Act) [Internet]. Available from: www.regjeringen.no/upload/HOD/HRA/Helseforskning/Helseforskningsloven%20-%20ENGELSK%20endelig%2029%2006%2009.pdf
36. 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 Apr 1;350(14):1414–21.
37. Young AF, Dobson AJ, Byles JE. Health services research using linked records: who consents and what is the gain? Aust N Z J Public Health. 2001 Oct;25(5):417–20.
38. 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 Dec 22;10(1):346.
39. 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 Nov 1;5(2):53–60.
40. 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 Mar 12;338(mar12 2):b866–b866.
41. FEAM releases Briefing Pack on data protection [Internet]. [cited 2014 Oct 24]. Available from: http://www.iamp-online.org/content/feam-releases-briefing-pack-data-protection
42. Nilstun T, Hermerén G. Human Tissue Samples and Ethics. Med Health Care Philos. 2006 Mar 1;9(1):81–6.
43. Tupasela A, Sihvo S, Snell K, Jallinoja P, Aro AR, Hemminki E. Attitudes towards biomedical use of tissue sample collections, consent, and biobanks among Finns. Scand J Public Health. 2010 Feb;38(1):46–52.
44. Emam KE, 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 Jun 9;11(1):454.
45. Bussey-Jones J, Garrett J, Henderson G, Moloney M, Blumenthal C, Corbie-Smith G. The role of race and trust in tissue/blood donation for genetic research. Genet Med. 2010 Feb;12(2):116–21.
46. Goldman J, Hudson Z. Virtually Exposed: Privacy And E-Health. Health Affairs (2000).
47. Emam KE, 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. J Am Med Inform Assoc. 2011 May 1;18(3):212–7.
48. Kowalczyk S, Shankar K. Data sharing in the sciences. Ann Rev Info Sci Tech. 2011 Jan 1;45(1):247–94.

49. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Medical Informatics and Decision Making. 2014 Jun 11;14(1):51.
50. Forrow L, Taylor WC, Arnold RM. Absolutely relative: how research results are summarized can affect treatment decisions. Am J Med. 1992 Feb;92(2):121–4.
51. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med [Internet]. 2005 Aug [cited 2015 May 13];2(8). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/
52. World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. Bull World Health Organ. 2001;79(4):373–4.
53. 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.
54. Nass SJ, Levit LA, Gostin LO, Rule I of M (US) C on HR and the P of HITHP. Privacy - Beyond the HIPAA Privacy Rule - NCBI Bookshelf [Internet]. 2009 [cited 2014 Sep 19]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK9572/def-item/gl25/
55. Heeney C, Hawkins N, de Vries J, Boddington P, Kaye J. Assessing the Privacy Risks of Data Sharing in Genomics. Public Health Genomics. 2011;14(1):17–25.
56. Sweeney L. Uniqueness of Simple Demographics in the U.S. Population. LIDAP-WP4 Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA: 2000. 1000;
57. Emam KE, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ. 2015 Mar 20;350:h1139.
58. Wu FT. Defining Privacy and Utility in Data Sets [Internet]. Rochester, NY: Social Science Research Network; 2012 Apr [cited 2015 May 14]. Report No.: ID 2031808. Available from: http://papers.ssrn.com/abstract=2031808
59. Verykios VS, Bertino E, Fovino IN, Provenza LP, Saygin Y, Theodoridis Y. State-of-the-art in Privacy Preserving Data Mining. SIGMOD Rec. 2004 Mar;33(1):50–7.
60. Bogdanov D, Kamm L, Laur S, Sokk V. Rmind: a tool for cryptographically secure statistical analysis [Internet]. 2014 [cited 2014 Oct 1]. Report No.: 512. Available from: http://eprint.iacr.org/2014/512
61. Yao AC. Protocols for Secure Computations. Proceedings of the 23rd Annual Symposium on Foundations of Computer Science [Internet]. Washington, DC, USA: IEEE Computer Society; 1982 [cited 2014 Oct 13]. p. 160–4. Available from: http://dx.doi.org/10.1109/SFCS.1982.88
62. Gentry C. Fully Homomorphic Encryption Using Ideal Lattices. Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing [Internet]. New York, NY, USA: ACM; 2009 [cited 2014 Oct 13]. p. 169–78. Available from: http://doi.acm.org/10.1145/1536414.1536440
63. Paillier P. Public-Key Cryptosystems Based on Composite Degree Residuosity Classes. In: Stern J, editor. Advances in Cryptology — EUROCRYPT '99 [Internet]. Springer Berlin Heidelberg; 1999 [cited 2014 Oct 13]. p. 223–38. Available from: http://link.springer.com/chapter/10.1007/3-540-48910-X_16
64. Chaum D, Crépeau C, Damgard I. Multiparty Unconditionally Secure Protocols. Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing [Internet]. New York, NY, USA: ACM; 1988 [cited 2014 Oct 13]. p. 11–9. Available from: http://doi.acm.org/10.1145/62212.62214
65. Bogdanov D, Talviste R. Survey on application models and usage scenarios for various SMC protocols. University of Tartu;
66. 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.
67. Goldreich O. Secure multi-party computation. Journal of Cryptology. 2000;
68. Benaloh JC. Secret Sharing Homomorphisms: Keeping Shares of a Secret Secret (Extended Abstract). In: Odlyzko AM, editor. Advances in Cryptology — CRYPTO' 86 [Internet]. Springer Berlin Heidelberg; 1987 [cited 2014 Oct 13]. p. 251–60. Available from: http://link.springer.com/chapter/10.1007/3-540-47721-7_19
69. 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 [Internet]. Springer New York; 2006 [cited 2014 Oct 13]. p. 237–61. Available from: http://link.springer.com/chapter/10.1007/0-387-35209-0_14
70. Clifton C, Kantarcioglu M, Vaidya J, Lin X, Zhu MY. Tools for Privacy Preserving Distributed Data Mining. SIGKDD Explor Newsl. 2002 Dec;4(2):28–34.

71. Du W, Atallah MJ. Privacy-preserving cooperative statistical analysis. Computer Security Applications Conference, 2001 ACSAC 2001 Proceedings 17th Annual. 2001. p. 102–10.
72. Du W, Zhan Z. Building Decision Tree Classifier on Private Data. Proceedings of the IEEE International Conference on Privacy, Security and Data Mining - Volume 14 [Internet]. Darlinghurst, Australia, Australia: Australian Computer Society, Inc.; 2002 [cited 2014 Oct 17]. p. 1–8. Available from: http://dl.acm.org/citation.cfm?id=850782.850784
73. Atallah MJ, Du W. Secure Multi-party Computational Geometry. In: Dehne F, Sack J-R, Tamassia R, editors. Algorithms and Data Structures [Internet]. Springer Berlin Heidelberg; 2001 [cited 2014 Oct 16]. p. 165–79. Available from: http://link.springer.com/chapter/10.1007/3-540-44634-6_16
74. Cheung DW, Han J, Ng VT, Fu AW, Fu Y. A fast distributed algorithm for mining association rules. , Fourth International Conference on Parallel and Distributed Information Systems, 1996. 1996. p. 31–42.
75. Zhang X, Sun H, Wen Q, Sha S. Study on the Technology of the Secure Computation in the Different Adversarial Models. In: Jin D, Lin S, editors. Advances in Computer Science, Intelligent System and Environment [Internet]. Springer Berlin Heidelberg; 2011 [cited 2014 Oct 23]. p. 473–8. Available from: http://link.springer.com/chapter/10.1007/978-3-642-23777-5_77
76. Vaidya J. Secure Multi-Party Computation [Internet]. Purdue University. Available from: http://www.pdfio.net/k-2482958.html
77. Moher D, Liberati A, Tetzlaff J, Altman DG, for the PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009 Jul 21;339(jul21 1):b2535–b2535.
78. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. Summit on Translat Bioinforma. 2010 Mar 1;2010:1–5.
79. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar 1;17(2):124–30.
80. Drake TA, Braun J, Marchevsky A, Kohane IS, Fletcher C, Chueh H, et al. A system for sharing routine surgical pathology specimens across institutions: the Shared Pathology Informatics Network. Hum Pathol. 2007 Aug;38(8):1212–25.
81. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, et al. A Survey of Informatics Platforms That Enable Distributed Comparative Effectiveness Research Using Multi-institutional Heterogenous Clinical Data. Medical Care July 2012 [Internet]. 2012 [cited 2014 Feb 10]; Available from: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=ovftn&AN=00005650-201207001-00012
82. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies. PLoS One [Internet]. 2013 Mar 7 [cited 2014 Oct 8];8(3). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3591385/
83. Natter MD, Quan J, Ortiz DM, Bousvaros A, Ilowite NT, Inman CJ, et al. An i2b2-based, generalizable, open source, self-scaling chronic disease registry. Journal of the American Medical Informatics Association. 2013 Jan 1;20(1):172–9.
84. SHRINE: High-level architecture [Internet]. Available from: https://open.med.harvard.edu/wiki/display/SHRINE/High-Level+Architecture
85. Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS) [Internet]. PCORnet. [cited 2014 Sep 16]. Available from: /clinical-data-research-networks/cdrn1-harvard-university-scihls/
86. System Architecture | PopMedNet [Internet]. [cited 2014 Sep 15]. Available from: http://www.popmednet.org/?page_id=31
87. Frequently Asked Questions | PopMedNet [Internet]. [cited 2014 Sep 22]. Available from: http://www.popmednet.org/?page_id=45#current_PMN_data_models
88. Klann JG, Buck MD, Brown J, Hadley M, Elmore R, Weber GM, et al. Query Health: standards-based, cross-platform population health surveillance. J Am Med Inform Assoc. 2014 Apr 3;amiajnl – 2014–002707.
89. Mini-Sentinel: Distributed Query Tools and Summary Tables [Internet]. Available from: http://mini-sentinel.org/data_activities/distributed_query_tool/details.aspx?ID=134
90. HMORN: HMO Research network [Internet]. Available from: http://www.hmoresearchnetwork.org/en/
91. Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. Health Aff (Millwood). 2014 Jul;33(7):1178–86.

92. 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 Aug;21(4):714–9.

93. SCANNER | Scalable National Network for Effectiveness Research: Semantic Interoperability [Internet]. Available from: http://scanner.ucsd.edu/about/semantic-interoperability

94. Voets D. EHR4CR. Initial EHR4CR architecture and interoperability framework specifications [Internet]. 2012. Available from: http://www.ehr4cr.eu/files/ExecutiveSummary/EHR4CR-ExecutiveSummaryD3_1.pdf

95. Lisa M. Schilling BMK. Scalable Architecture for Federated Translational Inquiries Network (SAFTINet) Technology Infrastructure for a Distributed Data Network. eGEMS. 2013;1(1):Article 11.

96. Youwen Z, Liusheng H, Wei Y, Xing Y. Efficient Collusion-Resisting Secure Sum Protocol. Chinese Journal of Electronics. 2011 Jul;20(3).

97. Chu C-T, Kim SK, Lin Y-A, Yu YY, Bradski G, Ng Y. A, et al. Map-Reduce for Machine Learning on Multicore. Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference [Internet]. [cited 2014 Oct 19]. Available from: http://cs.stanford.edu/people/ang/?portfolio=map-reduce-for-machine-learning-on-multicore

98. Shepard S, Kresman R, Dunning L. Data Mining and Collusion Resistance. proceedings of world congress on Engineering 2009. 2009 Jul 1;1(WCE 2009).

99. Beimel A. Secret-Sharing Schemes: A Survey. In: Chee YM, Guo Z, Ling S, Shao F, Tang Y, Wang H, et al., editors. Coding and Cryptology [Internet]. Springer Berlin Heidelberg; 2011 [cited 2015 May 7]. p. 11–46. Available from: http://link.springer.com/chapter/10.1007/978-3-642-20901-7_2

100. 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 [Internet]. Springer Berlin Heidelberg; 2011 [cited 2014 Oct 20]. p. 24–36. Available from: http://link.springer.com/chapter/10.1007/978-3-642-22890-2_3

101. 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.

102. 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 [Internet]. Anaheim, CA, USA: ACTA Press; 2007 [cited 2014 Oct 20]. p. 326–31. Available from: http://dl.acm.org/citation.cfm?id=1295581.1295634

103. Shuang Wang XJ. EXpectation Propagation LOgistic REgRession (EXPLORER): Distributed Privacy-Preserving Online Model Learning. Journal of biomedical informatics. 2013;

104. Hayes B. Alice and Bob in Cipherspace : A new form of encryption allows you to compute with data you cannot read. American Scientist. 2012 Oct;September-October 2012 Volume 100,(Number 5).

105. Kargupta H, Kargupta H, Das K, Das K, Liu K, Liu K. A game theoretic approach toward multi-party privacypreserving distributed data mining. In In Communication; 2007.

106. 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 [Internet]. New York, NY, USA: ACM; 2013 [cited 2014 Oct 23]. p. 4:1–4:6. Available from: http://doi.acm.org/10.1145/2541534.2541593

107. Bellika JG, Henriksen TS, Yigzaw KY. The Snow System: A Decentralized Medical Data Processing System - Springer. In: Fernández-Llatas C, García-Gómez JM, editors. Springer New York; 2015 [cited 2015 May 12]. Available from: http://link.springer.com/protocol/10.1007%2F978-1-4939-1985-7_7

108. Garde S, Knaup P, Hovenga E, Herd S. Towards Semantic Interoperability for Electronic Health Records: Domain Knowledge Governance for openEHR Archetypes. Methods of Information in Medicine [Internet]. 2007 [cited 2014 May 26]; Available from: http://www.schattauer.de/index.php?id=1214&doi=10.1160/ME5001

109. Clinical Knowledge Manager - Health Information Models - openEHR Wiki [Internet]. [cited 2014 May 29]. Available from: http://www.openehr.org/wiki/display/healthmod/Clinical+Knowledge+Manager

110. Nasjonal IKT: Clinical knowledge Manager [Internet]. Available from: http://arketyper.no/ckm/

111. Ma C, Frankel H, Beale T, Heard S. EHR query language (EQL)--a query language for archetype-based health records. Stud Health Technol Inform. 2007;129(Pt 1):397–401.

112. Archetype Query Language Description - Specifications - openEHR Wiki [Internet]. [cited 2014 May 26]. Available from: http://www.openehr.org/wiki/display/spec/Archetype+Query+Language+Description
113. Think!EHR Platform [Internet]. [cited 2014 Mar 10]. Available from: http://www.marand-thinkmed.com/thinkehr
114. Agile Methedology [Internet]. Available from: http://agilemethodology.org/
115. Manifesto for Agile Software Development [Internet]. 2006. Available from: http://agilemanifesto.org/
116. Mannaro K. Adopting Agile Methodologies in Distributed Software Development [Internet]. 2008 Feb. Available from: http://veprints.unica.it/53/1/mannaro_katiuscia.pdf
117. Scrum Methodology [Internet]. Available from: http://scrumreferencecard.com/scrum-reference-card/
118. Java SE - Downloads | Oracle Technology Network | Oracle [Internet]. [cited 2014 May 29]. Available from: http://www.oracle.com/technetwork/java/javase/downloads/index.html
119. Libraries – The XMPP Standards Foundation [Internet]. [cited 2014 May 26]. Available from: http://xmpp.org/xmpp-software/libraries/
120. Miller G, Williams L. Personas: Moving Beyond Role-Based Requirements Engineering. Microsoft and North Carolina State University. 2006;
121. Pruitt J, Grudin J. Personas: Practice and theory. In Proceedings of DUX 2003. 2003.
122. Robertson S, Robertson J. Mastering the Requirements Process: Getting Requirements Right. Addison-Wesley; 2012. 579 p.
123. Hailemichael MA, MARCO-RUIZ L, Bellika JG. (Accepted) Privacy-preserving Statistical Query and Processing on Distributed OpenEHR data. Medical Informatics Europe. 2015.
124. Kearns M. Efficient Noise-Tolerant Learning From Statistical Queries. Journal of the ACM. ACM Press; 1993. p. 392–401.
125. 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.
126. 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 [Internet]. New York, NY, USA: ACM; 2010 [cited 2013 Jun 7]. p. 987–98. Available from: http://doi.acm.org/10.1145/1807167.1807275
127. Duan Y. P4P: A Practical Framework for Privacy-Preserving Distributed Computation [Internet] [PhD thesis]. [Berkeley]: University of California; 2007. Available from: http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-165.html
128. Code of Conduct for information security [Internet]. 2013. Available from: https://ehelse.no/personvern-og-informasjonssikkerhet/norm-for-informasjonssikkerhet/documents-in-english
129. admin. XMPP Extensions – The XMPP Standards Foundation [Internet]. [cited 2015 May 14]. Available from: http://xmpp.org/xmpp-protocols/xmpp-extensions/
130. Grimaldi D, Provini F, Pierangeli G, Mazzella N, Zamboni G, Marchesini G, et al. Evidence of a diurnal thermogenic handicap in obesity. Chronobiol Int. 2014 Nov 21;32(2):299–302.
131. Landsberg L, Young JB, Leonard WR, Linsenmeier RA, Turek FW. Do the Obese Have Lower Body Temperatures? A New Look at a Forgotten Variable in Energy Balance. Trans Am Clin Climatol Assoc. 2009;120:287–95.
132. Archetype Editor 2.2.905 [Internet]. Available from: http://www.openehr.org/downloads/archetypeeditor/home
133. Template Designer [Internet]. Available from: http://www.openehr.org/downloads/modellingtools
134. Vatsalan D, Christen P, Verykios VS. A Taxonomy of Privacy-preserving Record Linkage Techniques. Inf Syst. 2013 Sep;38(6):946–69.
135. Lindell Y, Pinkas B. Secure Multiparty Computation for Privacy-Preserving Data Mining. Journal of Privacy and Confidentiality [Internet]. 2009 Apr 14;1(1). Available from: http://repository.cmu.edu/jpc/vol1/iss1/5

Appendices

Appendix A

Summary of literature review results selected for full-text reading.

No	Article Title	Author (Last, Initial)
1	MDPHnet: secure, distributed sharing of electronic health record data for public health surveillance, evaluation, and planning.	Vogel, J; et al.,2014
2	TMI! Ethical challenges in managing and using large patient data sets.	Juengst, ET, 2014
3	Leveraging the cloud for electronic health record access.	Coats, B; Acharya, S. 2014
4	Patient informed governance of distributed research networks: results and discussion from six patient focus groups.	Mamo, LA. et al.,2013
5	Privacy-preserving health data collection for preschool children.	Guan, S; Zhang, Y; Ji, Y. 2013
6	Acceptability and perceived barriers and facilitators to creating a national research register to enable 'direct to patient' enrolment into research: the Scottish Health Research Register (SHARE)	Grant A. et al., 2013
7	Behavioral health data in the electronic health record: privacy concerns slow sharing.	Greene, J, 2013
8	Electronic health records: new opportunities for clinical research.	Coorevits, P. et al., 2013
9	Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records.	Fernandes, AC. et al., 2013
110	The secure medical research workspace: an IT infrastructure to enable secure research on clinical data.	Shoffner, M; et al., 2013
11	Implementation of the CDC translational informatics platform--from genetic variants to the national Swedish Rheumatology Quality Register.	Abugessaisa, I. et al., 2013
12	SHRINE: enabling nationally scalable multi-site disease studies.	McMurry, AJ. et al., 2013
13	Managing protected health information in distributed research network environments: automated review to facilitate collaboration.	Bredfeldt, CE. et al., 2013
14	The Scottish Emergency Care Summary--an evaluation of a national shared record system aiming to improve patient care: technology report.	Morris, LM. et al., 2012
15	Towards an international electronic repository and virtual laboratory of open data and open-source software for telehealth research: comparison of international, Australian and Finnish privacy policies.	Suominen, H. 2012

16	A policy framework for public health uses of electronic health data.	McGraw, D; Rosati, K; Evans, B. 2012
17	Comparative-effectiveness research in distributed health data networks.	Toh, S. et al., 2011
18	Architecture of a consent management suite and integration into IHE-based Regional Health Information Networks.	Heinze, O. et al., 2011
19	Design and testing of an architecture for a national primary care chronic disease surveillance network in Canada.	Keshavjee, K. et al., 2011
20	Implementation of a secure and interoperable generic e-Health infrastructure for shared electronic health records based on IHE integration profiles.	Schabetsberger, T. et al., 2010
21	Sentinel e-health network on grid: developments and challenges.	De Vlieger, P. et al., 2010
22	Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2).	Murphy, SN. et al., 2010
23	Design of a national distributed health data network.	Maro, JC. et al., 2009
24	Perspectives for medical informatics. Reusing the electronic medical record for clinical research.	Prokosch, HU; Ganslandt, T. 2009
25	Towards a virtual anonymisation grid for unified access to remote clinical data.	Sinnott, R. et al., 2008
26	Multi-centric universal pseudonymisation for secondary use of the EHR.	Lo Iacono, L. 2007
27	Patient confidentiality in the research use of clinical medical databases.	Krishna, R; Kelleher, K; Stahlberg, E 2007
28	Toward secure distribution of electronic health records: quantitative feasibility study on secure E-mail systems for sharing patient records.	Gomi, Y; Nogawa, H; Tanaka, H. 2005
29	Evaluation of secure multi-party computation for reuse of distributed electronic health data	Yigzaw, K.Y.; Bellika, J.G. 2014
30	Healthcare information exchange system based on a Hybrid central/federated model	Ghane, K. 2014
31	Privacy preserving health data processing	Andersen, A.; Yigzaw, K.Y.; Karlsen, R. 2014
32	Privacy-Preserving Medical Reports Publishing for Cluster Analysis	Hmood, A.; Fung, B.C.M.; Iqbal, F. 2014
33	Disclosure risk of individuals: A k-anonymity study on health care data related to Indian population	Panackal, J.J.; Pillai, A.S.; Krishnachandran, V.N. 2014
34	Analyzing healthcare provider centric networks through secondary use of	Sauter, S.K.et al., 2014

	health claims data	
35	Privacy preservation, sharing and collection of patient records using cryptographic techniques for cross-clinical secondary analytics	Abdulrahman, H.; Poh, N.; Burnett, J. 2014
36	XDS-I Outsourcing Proxy: Ensuring Confidentiality While Preserving Interoperability	Ribeiro, L.S.et al., 2014
37	The Role of Inference in the Anonymization of Medical Records	Zigomitros, A.; Solanas, A.; Patsakis, C. 2014
38	Security issues for data sharing and service interoperability in eHealth systems: The Nu.Sa. test bed	Frontoni, Eet al., 2014
39	Electronic Personal Health Records and the Question of Privacy	Li, J. 2013
40	OpenEHR-based pervasive health information system for primary care: First Brazilian experience for public care	Bacelar-Silva, G.M.et al., 2013
41	e-Enabling International Cancer Research: Lessons Being Learnt in the ENS@T-CANCER Project	Stell, A.; Sinnott, R. 2013
42	Gateway to the Cloud - Case Study: A Privacy-Aware Environment for Electronic Health Records Research	Smith, R.et al., 2013
43	Different Perception and Attitude toward Medical Data That Including Protected Health Information in Clinical Research	Mi Jung Rho; Kwang Soo Jang; In Young Choi 2013
44	Secure Electronic Health Record Exchange: Achieving the Meaningful Use Objectives	Acharya, S. et al., 2013
45	A Security Model for Distributed Collaborative Environments in the Healthcare	Viana Lopes Araujo, R.; Silva, F.J.S.E. 2013
46	An implementation of secure multi-party computations to preserve privacy when processing EMR data	Andersen, A. 2013
47	A framework for privacy-preserving healthcare data sharing	Lei Chen; Ji-Jiang Yang; Qing Wang; Yu Niu 2012
48	Privacy Preserving in Data Mining Using Hybrid Approach	Lohiya, S.; Ragha, L. 2012
49	A Hierarchical Framework for Secure and Scalable EHR Sharing and Access Control in Multi-cloud	Jie Huang; Sharaf, M.; Chin-Tser Huang 2012
50	Distributed Privacy Preserving Decision Support System for Predicting Hospitalization Risk in Hospitals with Insufficient Data	Mathew, G.; Obradovic, Z. 2012
51	Privacy-Preserving Data Publishing for Free Text Chinese Electronic Medical Records	Lei Chen; Ji-Jiang Yang; Qing Wang 2012
52	HCPP: Cryptography Based Secure EHR System for Patient Privacy and Emergency Healthcare	Jinyuan Sun; Xiaoyan Zhu; Chi Zhang; Yuguang Fang

		2011
53	Cross-Domain Data Sharing in Distributed Electronic Health Record Systems	Jinyuan Sun; Yuguang Fang 2010
54	Security Design for Electronic Medical Record Sharing System	Qingzhang Chen; Zhehu Wang; Wangqiao Zhang 2010
55	Digital Health Care (DHC) Network and IT Infrastructure Solutions	Wei Liu 2010
56	Designing Privacy for Scalable Electronic Healthcare Linkage	Stell, A.et al., 2009
57	Privacy-preserving electronic health record linkage using pseudonym identifiers	Alhaqbani, B.; Fidge, C. 2008
58	Security Oriented e-Infrastructures Supporting Neurological Research and Clinical Trials	Stell, A. et al., 2007
59	Improving outcomes with interoperable EHRs and secure global health information infrastructure	Kun, L.et al., 2007
60	Collaborative Support for Medical Data Mining in Telemedicine	Wong Kok Seng; Besar, R.B.; Abas, F.S. 2006
61	Publishing data from electronic health records while preserving privacy	Aris Gkoulalas-Divanis, Grigorios Loukides, Jimeng Sun 2014
62	A data recipient centered de-identification method to retain statistical attributes	Gal, S.T. et al., 2014
63	A flexible approach to distributed data anonymization	Kohlmayer F. et al., 2014
64	A new tool for sharing and querying of clinical documents modeled using HL7 Version 3 standard	Slavov V. et al., 2013
65	Trustworthy reuse of health data: A transnational perspective	A. Geissbuhler et al., 2013
66	Perspectives of Australian adults about protecting the privacy of their health information in statistical databases	Tatiana King, Ljiljana Brankovic, Patricia Gillard 2012
67	Scope, rationale and design of an infrastructure for the study of physical and psychosocial outcomes in cancer survivorship cohorts	van de Poll-Franse, V.L. et al., 2012
68	A methodology for the pseudonymization of medical data	Thomas Neubauer, Johannes Heurix 2011
69	Views on health information sharing and privacy from primary care practices using electronic medical record	Perera G. et al., 2011
70	eHealth in Belgium, a new "secure" federal network: Role of patients,	Francis Roger France,2010

	health professions and social security services	
71	Towards a Virtual Research Environment for International Adrenal Cancer Research	Richard O. Sinnott, Anthony J. Stell, 2011
72	A method to implement fine-grained access control for personal health records through standard relational database queries	Sujansky, V.W. et al., 2010
73	Strategies for health data exchange for secondary, cross-institutional clinical research	Elger, S.B. et al., 2010
74	A Globally Optimal k-Anonymity Method for the De-Identification of Health Data	Emam, El Kh. et al., 2009
75	Properties of a federated epidemiology query system	Bellika, J.G. et al., 2007
76	An ‘Honest Broker’ mechanism to maintain privacy for patient care and academic medical research	Boyd, D.A. et al., 2007
77	An e-consent-based shared EHR system architecture for integrated healthcare networks	Bergmann, J. et al., 2007
78	Toward a National Framework for the Secondary Use of Health Data	Safran, Ch. et al., 2007
79	Internet Based Multi-Institutional Clinical Research: A Convenient and Secure Option	Lallas, D.C. et al., 2004
80	Towards privacy-preserving computing on distributed electronic health record data	Yigzaw Y.K. et al., 2013
81	On the challenges of balancing privacy and utility of open health data	Guttmann C. et al., 2013
82	An Ensemble Topic Model for Sharing Healthcare Data and Predicting Disease Risk	Andrew K. Rider, Nitesh V. Chawla 2013
83	The application of differential privacy to health data	Fida Kamal Dankar, Khaled El Emam 2012
84	A software tool for large-scale sharing and querying of clinical documents modeled using HL7 version 3 standard	Rao, R.P. et al., 2012
85	Proceedings of the 2011 workshop on Data mining for medicine and healthcare	Chawla N. et al., 2011
86	Privacy-enhanced management of ubiquitous health monitoring data	George Drosatos, Pavlos S. Efraimidis 2011
87	An application architecture to facilitate multi-site clinical trial collaboration in the cloud	Jonathan Sharp 2011
88	Privacy preserving EHR system using attribute-based infrastructure	Narayan Sh. et al., 2010
89	ρ-uncertainty: inference-proof transaction anonymization	Cao J., et al., 2010

90	Patient-centric authorization framework for sharing electronic health records	Jin J.et al., 2009

Appendix B

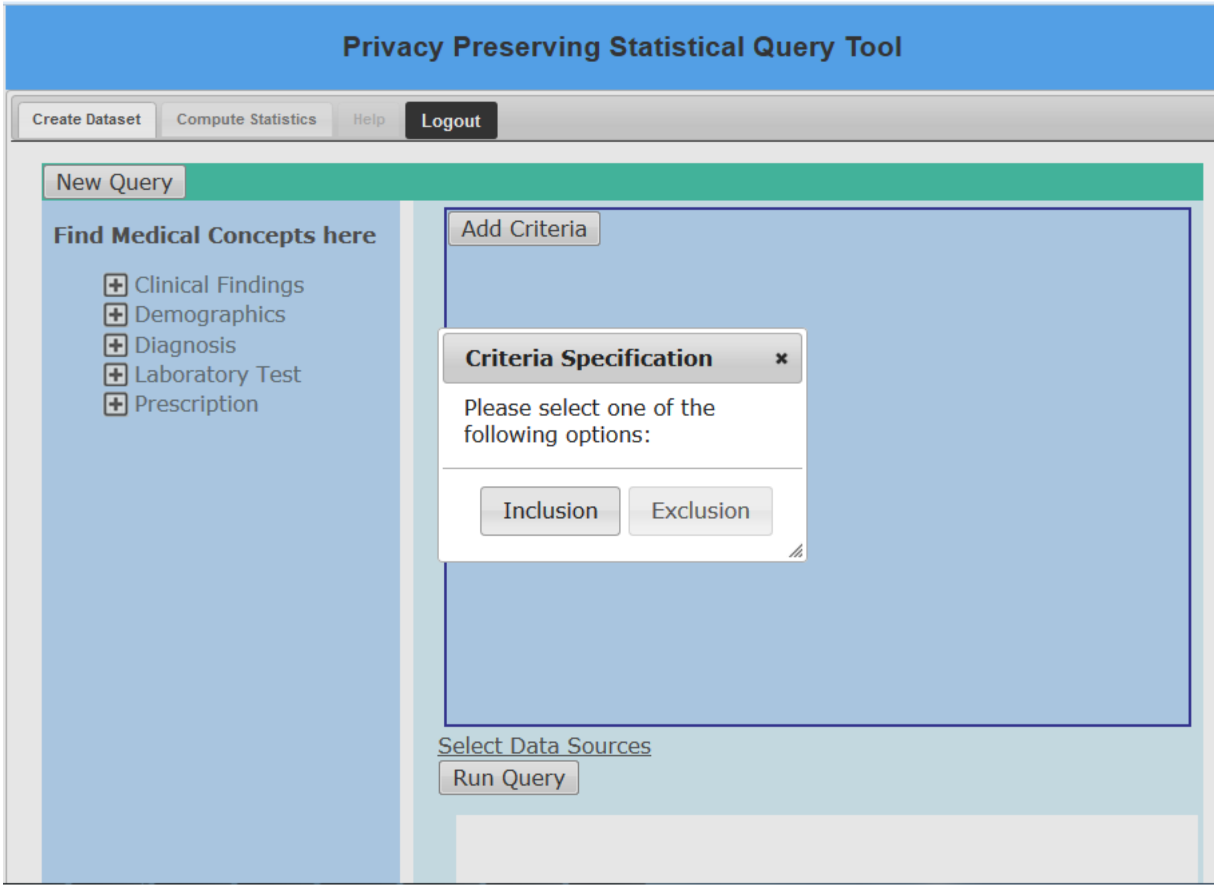
Statistical functions decomposed into summation form

#	Function	Formula	Description
1	Count	$\boldsymbol{Count}([x]) = n$	Total count of variable *x*
2	Average	$\boldsymbol{mean} = \frac{1}{n}(\sum_{i=1}^{n}[x_i])$	Where x_i is an individual value and ***n*** is count
3	Chi Square test	$\boldsymbol{chi^2} = \frac{\sum_{i=1}^{n}([O - \boldsymbol{E}])^2}{E}$	Where O is Observed frequency and E is Expected frequency
4	Var	$\boldsymbol{var} = \frac{(\sum_{i=1}^{n}([x_i] - \boldsymbol{mean}([x]))^2)}{\boldsymbol{n}}$	Where $\boldsymbol{mean}(x)$ is mean of variable *x* and ***n*** is count
5	Sdev	$\boldsymbol{sdev}([x]) = \sqrt{\boldsymbol{var}([x])}$	Where $\boldsymbol{var}([x])$ is variance of *x*
6	Stand.err	$\boldsymbol{mean} = \frac{\boldsymbol{sdev}([x])}{\sqrt{n}}$	Where $\boldsymbol{sdev}(x)$ is the standard deviation of *x* and ***n*** is count
7	Covar	$\boldsymbol{Covar} = \frac{\sum_{i=1}^{n}([x_i] - \boldsymbol{mean}([x]))([y_i] - \boldsymbol{mean}([y]))}{\boldsymbol{n}}$	Where $\boldsymbol{mean}(x)$ is mean of variable *x*, $\boldsymbol{mean}(y)$ is mean of *y* and ***n*** is the count
8	F-test	$F\ value = \frac{\boldsymbol{Var}(x)}{\boldsymbol{Var}(y)}$	Where $\boldsymbol{Var}(x)$ is variance of variable *x* and $\boldsymbol{Var}(y)$ is variance of *y*.
10	Kurt	$Kurtosis = \frac{\sum_{i=1}^{n}([x_i] - \boldsymbol{mean}([x]))}{(\boldsymbol{n} - 1)\boldsymbol{sdev}(x)}$	Where *x* is an individual value at *i* and $\boldsymbol{mean}([x])$ is the average, ***n*** is count and $\boldsymbol{sdev}(x)$ is the standard deviation

11	Pearson's r/ correlation	$\boldsymbol{r} = \frac{\boldsymbol{Covar}([x][y])}{\sqrt{\boldsymbol{sdev}([x])\boldsymbol{sdev}([y])}}$	Where $\boldsymbol{Covar}([x][y])$ is covariance of variable x and y, $\boldsymbol{sdev}(x)$ and $\boldsymbol{sdev}(y)$ are the standard deviation of x and y respectively
12	Skew	$Skew = \frac{1}{\boldsymbol{n}} \frac{\sum_{i=1}^{n}([x_i] - \boldsymbol{mean}([x]))^3}{\boldsymbol{sdev}(x)^3}$	Where $\boldsymbol{mean}([x])$ is the average, $\boldsymbol{n}$ is count and $\boldsymbol{sdev}(x)$ is the standard deviation of x
13	Linear regression	$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}.\boldsymbol{x}$ $\boldsymbol{\beta_1} = \frac{\boldsymbol{Covar}([\boldsymbol{x}][\boldsymbol{y}])}{\boldsymbol{var}([\boldsymbol{x}])}$ $\boldsymbol{\beta_0} = \boldsymbol{mean}([y]) - \boldsymbol{\beta_1}.\boldsymbol{mean}([x])$	Where $\boldsymbol{y}$ is dependent variable, $\boldsymbol{x}$ is independent variable, $\boldsymbol{\beta_0}$ is intercept, $\boldsymbol{\beta_1}$ is slope, $\boldsymbol{Covar}([x][y])$ is covariance of x and y, $\boldsymbol{var}(x)$ is the variance of x, $\boldsymbol{mean}(x)$ and $\boldsymbol{mean}(y)$ are mean of x and y respectively
14	T-test	$T - test = \frac{\boldsymbol{mean}([x]) - \boldsymbol{mean}([y])}{\sqrt{\frac{\boldsymbol{sdev}([x]) + \boldsymbol{sdev}([y])}{\boldsymbol{n}}}}$	Where $\boldsymbol{sdev}(x)$ is the standard deviation of x, $\boldsymbol{n}$ is count, $\boldsymbol{mean}(x)$ and $\boldsymbol{mean}(y)$ are mean of x and y respectively

Appendix C1

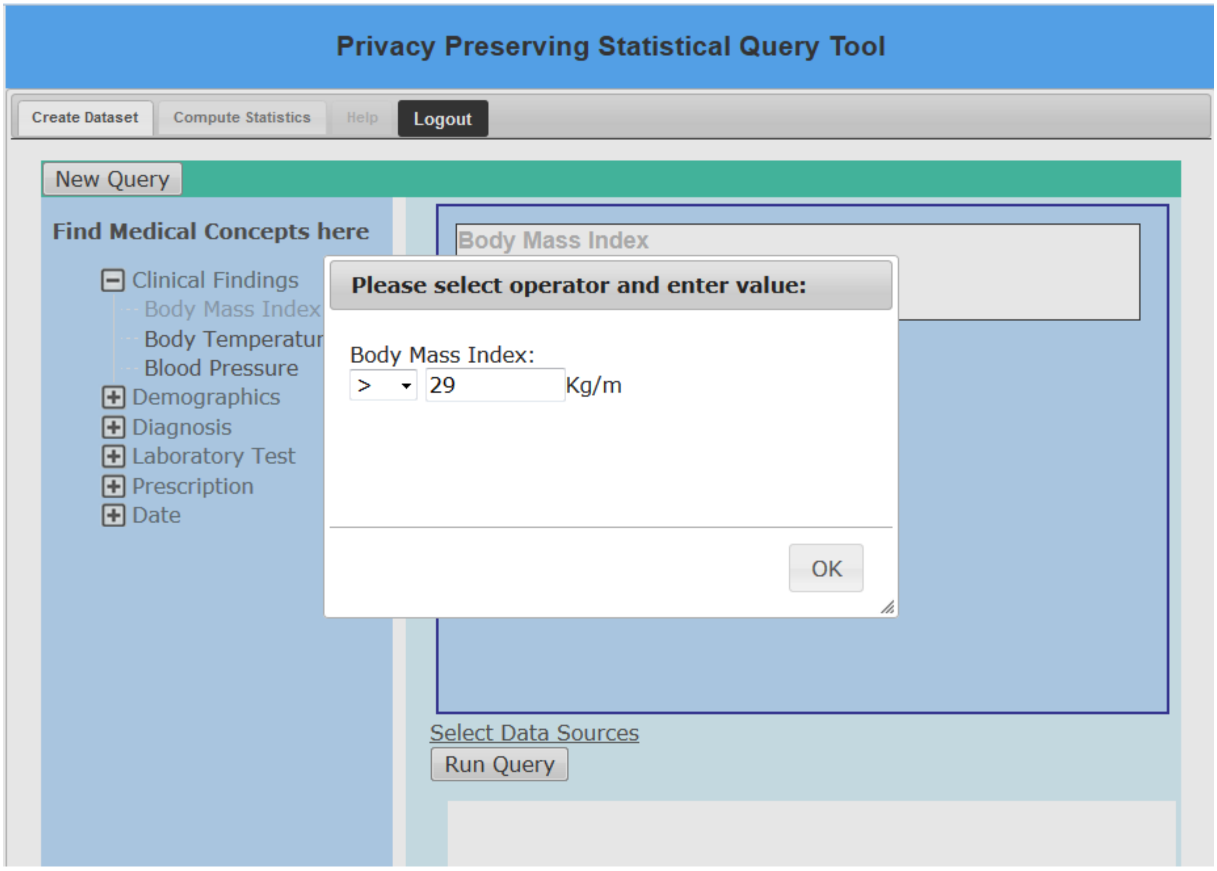
A dialog box to specify whether a criterion is *Inclusion* or *Exclusion*



Please select one of the following options:

Appendix C2

A dialog box to select an operator and enter the values of the selected parameters

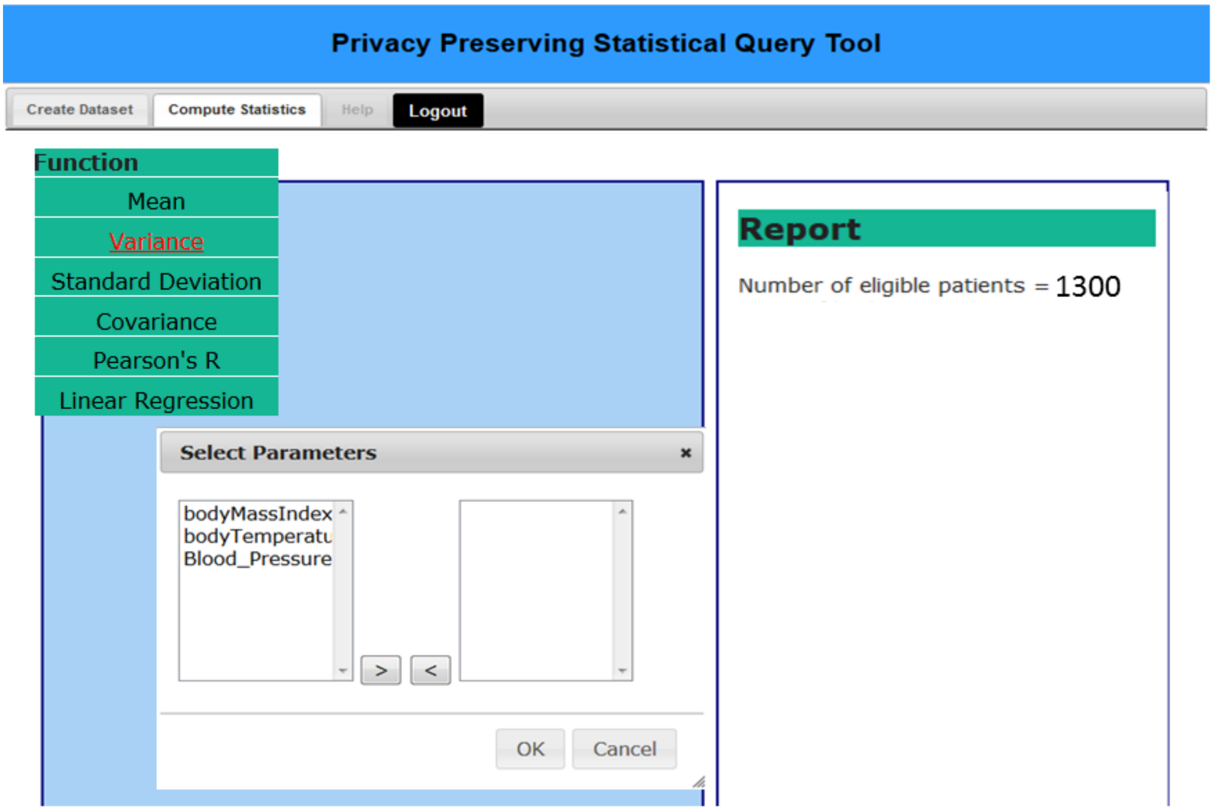


Please select operator and enter value:

Body Mass Index:
> 29 Kg/m

Appendix C3

A dialog box to select the parameters for the selected statistical function



Number of eligible patients = 1300