Privacy-Preserving Data Integration in Public Health Surveillance

Jun Hu

Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

In partial fulfillment of the requirements

For the PhD degree in Computer Science

School of Information Technology and Engineering

Faculty of Engineering

University of Ottawa

Abstract

With widespread use of the Internet, data is often shared between organizations in B2B health care networks. Integrating data across all sources in a health care network would be useful to public health surveillance and provide a complete view of how the overall network is performing. Because of the lack of standardization for a common data model across organizations, matching identities between different locations in order to link and aggregate records is difficult. Moreover, privacy legislation controls the use of personal information, and health care data is very sensitive in nature so the protection of data privacy and prevention of personal health information leaks is more important than ever. Throughout the process of integrating data sets from different organizations, consent (explicitly or implicitly) and/or permission to use must be in place, data sets must be de-identified, and identity must be protected. Furthermore, one must ensure that combining data sets from different data sources into a single consolidated data set does not create data that may be potentially re-identified even when only summary data records are created.

In this thesis, we propose new privacy preserving data integration protocols for public health surveillance, identify a set of privacy preserving data integration patterns, and propose a supporting framework that combines a methodology and architecture with which to implement these protocols in practice. Our work is validated with two real world case studies that were developed in partnership with two different public health surveillance organizations.

Acknowledgements

I would like first and foremost to thank my supervisor, Dr. Liam Peyton. Without his great support, incredible encouragement and patient guidance throughout my research this work would have not been possible to complete. I also wish to express my most sincere gratitude to Dr. Khaled El Emam for his breadth of knowledge, great ideas and kind advice that has been an invaluable contribution to this thesis.

I am grateful to Dr. Herna L Viktor, Dr. Michael Weiss and Dr. Ashraf Aboulnaga (University of Waterloo) for accepting to be members of my committee. I also owe many thanks to Dr. Daniel Amyot for his encouragement and support.

Special thanks are dedicated to my family for their love, unconditional support and encouragement throughout my study and research in past years.

I also would like to thank Saeed Samet from CHEO Research Institute for his excellent suggestions on secure protocols, as well as Tom Wong and Gayatri Jayaraman from the Public Health Agency of Canada for their support on the HPV case study.

This work was supported by a postgraduate research grant from the National Sciences and Engineering Research Council of Canada (NSERC), by a collaborative health research project grant from CIHR and NSERC (Canada) on "Performance management at the point of care: secure data delivery to drive clinical decision making processes for hospital quality control", and by the University of Ottawa through an entrance scholarship.

Table of Contents

List of Figures

List of Tables

List of Acronyms

Acronym	Definition
AR	IBM DB2 Anonymous Resolution
B2B	Business-to-Business
BI	Business Intelligence
CHEO	Children's Hospital of Eastern Ontario
CoT	Circle of Trust
CRISP-DM	Cross-Industry Standard Process for Data Mining
EHR	Electronic Health Records
EMR	Electronic Medical Record
ETL	Extract Transform Load
GI	Gastro-Intestinal
HL7	Health Level Seven
HPV	Human Papilloma Virus
ILI	Influenza Like Illness
ISDS	International Society for Disease Surveillance
MPI	Master Patient Index
OLAP	On-Line Analytic Processing
PAP Test	Papanicolaou Smear
PHI	Personal Health Information
PHR	Personal Health Record
PPRL	Privacy Preserving Record Linkage
RBAC	Role-based Access Control
SaaS	Software as a Service
SOA	Service Oriented Architecture
TBAC	Team-based Access Control

Chapter 1. Introduction

A framework for effective and secure integration of appropriate information in a healthcare network on a real-time, continuous basis would be useful for public health surveillance reporting. This should be feasible by leveraging Internet technologies. However, different organizations hold their data in different formats and matching identities between different locations in order to link and aggregate records is difficult. Moreover, privacy legislation (HIPAA, 1996, PIPEDA, 2000, PHIPA 2004) controls the use of personal information; and health care data is very sensitive in nature so the protection of data privacy and prevention of personal health information leaks is more important than ever. Consent must be obtained, data sets must be de-identified, and identity must be protected throughout the process of integrating data sets. Furthermore, one must ensure that combining data sets from different data sources into a single consolidated data set does not create data that may be potentially re-identified (El Emam 2006) even when only summary data records are created.

In this thesis, we propose three new privacy preserving data integration protocols for public health surveillance; and we identify a set of privacy preserving data integration patterns classified based on essential principles for privacy protected data integration; as well, we develop a supporting framework combining a methodology and architecture to adapt these protocols into practice. Two case studies developed in collaboration with public health surveillance organizations to address real situations are used to illustrate and evaluate the contributions of the thesis.

1.1. Problem Statement

Public health surveillance often requires collecting data from several organizations. However, once an organization submits its data, it loses control of the data, and the privacy of its data is fully dependent on the collector. Privacy concerns will limit the willingness of data custodians to share data.

In most public health surveillance systems, patient level data is not needed, just aggregated counts. For example, in disease surveillance only counts of patients presenting with the particular syndrome under surveillance are needed; individual patient level data is not. However, to detect meaningful trends or spikes, demographic information is helpful to detect and localize epidemics within geographically or demographically defined sub-populations. This type of data, though, is susceptible to re-identification techniques (El Emam et al, 2006). As well, concerns extend beyond preserving patient privacy, but also include protecting the confidentiality and reputation of the data provider. Individual doctors may not want counts to be linked to their practice.

In other circumstances, public health surveillance requires linking patient level data across data sources from different organizations and generating the statistical reports for analysis. Very often, the medical researcher wants to learn the relationship between several attributes based on an integrated, composite view of a single patient whose data is located in multiple and disparate data sources. However, the distributed data sources are not allowed to share data based on patient identity. Moreover when linking data from different data sources, the combination of data may be potentially re-identified (El Emam et al, 2006).

This thesis aims to address these privacy issues in data integration frameworks and protocols for public health surveillance. Overall, we conducted our research to answer the following research questions:

1. What are the key issues for privacy-preserving data integration in health care?
2. What are the gaps between the requirements and existing solutions?
3. How can we leverage modern technologies and techniques to improve existing data integration solutions?
4. How can we protect personal health information (PHI) and prevent PHI leaks while enabling data integration to support public health surveillance?

1.2. Objectives

The main objective of this thesis is to develop new privacy preserving data integration protocols to solve privacy problems for public health surveillance, as well as leverage modern, intelligent methods and techniques to improve existing Internet data integration methodology and architecture for supporting these protocols in practice. The protocols and supporting framework are illustrated and evaluated with two case studies that have been developed in collaboration with public health organizations to address issues faced in current practice.

1.3. Contributions

The main contributes of this thesis are:

1. The development of three new protocols for privacy-preserving data integration for public surveillance:

    - A k-key holder protocol for aggregation that protects the identity of data providers, (in fact, it is the first protocol in existing literature to do so) ;
    - A federated pseudonymous linking protocol that protects the identity of patients;
    - A secure multi-party computation protocol that links patients anonymously while securely computing statistics.

2. A set of privacy preserving data integration patterns systematically classified based on essential principles for privacy protected data integration. The patterns bundle protocols and techniques to address different types and levels of PHI protection (anonymized vs federated pseudonyms) for different types of data integration (aggregated vs identity linking) according to the level of acceptable risk and trust involved.
3. A framework to systematically manage privacy preserving data integration for public health surveillance that combines methodology, architectural patterns and a wide spectrum of protocols which can be tailored to fit specific scenarios.
4. An architecture for continuous privacy-preserving data integration within B2B health care networks that separates identity information from health information with third party services for managing identity (Identity Provider) managing data integration and publishing (Data Integration Service); and for managing

secure aggregation and computation (Aggregators and Key Holder Committee); as well as registering data integration data sets and managing inter-organizational agreements and access control in a B2B health care network (Dataset Registry).

5. A methodology for organizing and managing continuous privacy-preserving health surveillance in a B2B network that revises and extends the CRISP-DM (Shearer, 2000) methodology standard for data mining adopted by the European Union by grounding it in an appropriate architecture, identifying the steps in the process where privacy preserving must be addressed, and identifying the places where the appropriate techniques and technology are used within the context of the architecture.

The following papers, related to this thesis have been published or submitted for publication as well:

1. K. El Emam, **J. Hu**, S. Samet, L. Gaudette, L. Peyton, C. Earle, G. Layaraman, T. Wong (under review), "Secure Computation across Health Data Registries", submitted to Journal of the American Medical Informatics Association (JAMIA), 2010.
2. K. El Emam, **J. Hu**, J. Mercer, L. Peyton, M. Kantarcioglu, B. Malin, D. Buckeridge, S. Samet, C. Earle (under second review), "A Secure Protocol for Protecting the Identity of Providers When Disclosing Data for Disease Surveillance." In the Journal of the American Medical Informatics Association, 18:212-217, DOI: 10.1136/amiajnl-2011-000100, 2011.

3. **J. Hu**, L. Peyton, K. El Emam, "A Systematic Approach to PHI Leak Prevention in Continuous Health Care Data Integration", Proceedings of the workshop on Intelligent Methods for Protecting Privacy and Confidentiality in Data, AI2010, pp.5-11, 2010.
4. L. Peyton, **J. Hu**, "Identity Management and Audit Trail Support for Privacy Protection in E-Health Networks", Certification and Security in Health-Related Web Applications: Concepts and Solutions, A. Chryssanthou, I. Apostolakis, & I. Varlamis (Eds.), ISBN13: 9781616928957. Pages 160-173. IGI Global, Hershey, PA, USA. 2010.
5. **J. Hu**, L. Peyton, "A Framework for Privacy Assurance and Ubiquitous Knowledge Discovery in Health 2.0 Data Mashups", Ubiquitous Health and Medical Informatics, S. Mohammed, J. Fiaidhi (Eds.), ISBN13: 9781615207770. Pages 64-83. IGI Global, Hershey, PA, USA. 2010.
6. L. Peyton, **J. Hu**, "Federated Identity Management to Link and Protect Healthcare Data", International Journal of Electronic Business (IJEB). 8(3). Inderscience Publishers. 2010.
7. **J. Hu**, L. Peyton, "Integrating Identity Management with Federated Healthcare Data Models", 4th International MCeTech Conference on eTechnologies, Ottawa, Canada, May, 2009. LNBIP 26, Springer, pp 100-112.
8. **J. Hu**, L. Peyton, C. Turner, H. Bishay, "A Model of Trusted Data Collection for Knowledge Discovery in B2B Networks", Proceedings of the 2008 IEEE International MCETECH Conference on e-Technologies, pp. 60-69, 2008.

9. L. Peyton, **J. Hu**, “Knowledge Discovery in a Circle of Trust”, Data Mining VIII: Data, Text and Web Mining and their Business Applications, A. Zanasi, C. Brebbia, N. Ebecken (Eds.), WIT Press, Billerica, MA, USA, pp 235-244, 2007.
10. L. Peyton, **J. Hu**, B. Zhan, "Addressing Trust and Privacy in Telemedicine ", The Symposium on E-Commerce and E-Business in China (SEEC) at the Ninth International Conference on E-Commerce (ICEC 2007), 2007.
11. L. Peyton, **J. Hu**, C. Doshi, P. Seguin ,"Addressing Privacy in a Federated Identity Management Network for E-Health", 8th World Congress on the Management of eBusiness, Toronto, July, 2007.

1.4. Research Methodology

We employ a design-oriented research methodology (Hevner et al, 2004) to conduct our research. First, we identify the problems and requirements we would like to address by analyzing a few representative scenarios and reviewing the literature. Next, we identify the gaps between the requirements and existing approaches. Then we develop our protocols and framework iteratively to address the requirements and identified gaps. Finally we implement our protocols and framework to evaluate it against the representative case studies in comparison with other approaches. The evaluation is intended to show the utility of our approach for addressing the identified gaps. It is not intended to be an empirical evaluation (Hevner et al, 2004; March et al, 1995).

1.5. Thesis Organization

Chapter Two describes the background information of public health surveillance and a literature review of related work.

Chapter Three defines the evaluation criteria and conducts a gap analysis of data integration for public health surveillance based on three example scenarios and existing approaches.

Chapter Four describes three new protocols for data integration in public health surveillance.

Chapter Five illustrates an overview of our proposed framework for privacy preserving data integration in health surveillance.

Chapter Six describes a systematic set of patterns for privacy protecting data integration that bundles protocols and techniques to address different types and levels of PHI protection (anonoymized versus federated pseudonyms) for different types of data integration (aggregated vs identity linking) according to the level of acceptable risk and trust involved.

Chapter Seven describes in detail our two case studies which were developed in collaboration with public health surveillance organizations. One uses a secure computation protocol to facilitate aggregation of counts for disease surveillance to illustrate our approach to protect the identity of provider. The other one uses both federated pseudonyms and secure multi-party computation protocol to facilitate secure linking and compute aggregates for nationwide HPV surveillance.

Chapter Eight presents an evaluation of our approaches and discuss how it compares to other approaches.

Chapter Nine summarizes the conclusions of the thesis and identifies future work that can follow on from the thesis.

Chapter 2. Background and Related Work

This chapter presents background information and a literature review for privacy-preserving data integration in public health surveillance. The related work covers existing data integration methodology, architectures, record linkage, security and privacy mechanisms that can be used in data integration protocols.

2.1. Public Health Surveillance

Public health surveillance is "the ongoing, systematic collection, analysis and interpretation of data essential to the planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those responsible for prevention and control." (Public Health Surveillance, 2010). Data integration is a critical process for public health surveillance. It is fundamentally about querying across different data sources from physicians, hospitals, surveys, registries and other providers in healthcare network. This section discusses the background information in public health surveillance.

2.1.1 Ethics and Legislations in Health Care

Health care data is highly regulated because of its highly sensitive nature. In Canada, the federal Personal Information Protection and Electronic Documents act (PIPEDA, 2000) regulates how personal information can be shared and under what circumstances. In Ontario, the Personal Health Information Privacy Act (PHIPA, 2004) guides how health information custodians are to handle personal health information. In

the United States, the Health Insurance Portability and Accountability Act (HIPAA) (HIPPA, 1996) specifies the privacy rules and guideline for protecting health information.

There are also organizational guidelines for access protected health information. In most situations, data will not be provided until the entire surveillance process is reviewed. Ethical review boards, privacy review boards and technical review boards all provide the guidance to access data for health research and study. For example, The Ottawa Hospital has established an approval process for access of patient data (OHREB, 2010). Any research project involving human subjects must be reviewed and approved by the Ottawa Hospital Research Ethics Board before work is started. In Manitoba, three approvals are required to access data systems in the custody of the Manitoba Centre for Health Policy (MCHP, 2010). First a research proposal is reviewed for feasibility by MCHP. Second, approval is obtained from the data provider and the health research ethics board (HREB, 2010). Third, a research agreement is signed by the primary investigator. If the aim is to integrate data from several different organizations, then there are heightened concerns about the potential for re-identification that could compromise patient privacy (El Emam et al 2006).

2.1.2 Public Health Information Management

Public health surveillance relies largely on integrated, accurate and timely health information that is widely distributed in B2B health care networks. However, it is still not unusual to find public health surveillance that relies on voluntary reporting by fax or one-time agreements that deliver data on files. Typically, health care systems are hierarchically structured on a national basis from national health authority, provincial

health authorities, and local health authorities to clinics, lab and hospitals, to family doctors. In addition, disease registries are often built separately in different jurisdictions, even within the same country. Because of standardization issue and privacy concerns, it is really a challenge to integrate data across a health care network and to obtain a complete view of performance for a health care network.

Electronic Records

Electronic health records (EHR), electronic medical records (EMR) and personal health records (PHR) are increasingly the main basis for health information management. EHR is a longitudinal electronic record of patient health information, which is generated and maintained within a single hospital or clinic or a group of cooperating hospitals and clinics. An EMR is an electronic record of patient health information which is typically maintained by a single physician. PHR is generally defined as an EHR controlled by an individual patient. Both Google (Google Health, 2009) and Microsoft Healthvault (HealthVault, 2009) enable patients to maintain their PHR and make them shared among health care providers when needed.

Patient's Consent

In most jurisdictions (HIPAA, 1996, PIPEDA, 2000, PHIPA 2004), consent must be obtained when collecting patient health data for public health surveillance. Explicit consent requires that patients give their signed permission. Their data cannot be collected if they refuse to give permission. Implicit consent is based on the fact that health surveillance is arguably part of some healthcare service that is provided to patients. If a patient requests this service, their consent is implied. General consent, or blanket consent

without specific purpose can be used at the time of medical treatment. If the data is de-identified, aggregate data can be shared under certain circumstances without consent.

De-identification

Privacy legislation requires that person health information (PHI) is de-identified when it is disclosed for public health surveillance or other secondary use. In (HIPAA 1996), two methods are available for de-identifying protected health information: removing 18 specific identifiers and statistical methods. In addition, many different technologies and techniques can be used for de-identification., A high level overview of these de-identification techniques is provided in (El Emam et al 2009). It also gives some guidance on how to use de-identification techniques such as randomization, irreversible and reversible coding, heuristics and analytics.

Master Patient Index

A Master Patient Index (MPI) is often used to manage health records within a single institution. A MPI is a database that contains a unique identifier for each patient which can be used to lookup and link health records for a patient from a variety of sources, each of which may have their own way of identifying patients. An enterprise Master Patient Index that links records is commonly created within the context of a single organization like a hospital (Adragna, 1998). Some countries have created initiatives to build a national Master Patient Index (Tan, 1995; Neame & Olson, 1996; Walker, 1998) to support integrated health services. One of the issues in leveraging a Master Patient Index for public health surveillance is to ensure privacy protection when sharing and analyzing data. When more data is linked together, there is a greater potential a risk of re-

identify a patient even if the identities of the patient are removed. (Li and Shaw, 2005) presented an attribute analysis framework to de-identify personal health information for data mining. (El Emam et al, 2006) has evaluated how well de-identification protects privacy by "re-identifying" individuals based on combining "de-identified" data from distributed sources.

2.1.3 Standards and Tools for Public Health Surveillance

Public health surveillance collects and analyses data through knowledge discovery (MacQueen 1967, Fayyad et al 1996), data mining processes, and other business intelligence (BI) tools. Common functions of knowledge discovery and BI technologies are reporting, online analytical processing (OLAP) (Thomsen, 2002; Ledbetter and Morgen, 2001), analytics, data mining, business performance management, text mining (Hearst, 1999), statistical techniques (Jun et al, 1999) and predictive analytics. Often BI applications gather distributed, adhoc and unstructured data from various data sources into centralized data warehouses (Kimball et al,1998; Inmon, 2005), data marts, registries or portals for analysis and to support better business decision making and overall business performance management.

There are a number of standards that facilitate collecting and sharing of data in the health care industry. ETL tools are commonly used to extract, transform and load data into standardized datasets. Health Level Seven (HL7, 2010) is an ANSI-accredited Standard Developing Organization (SDO) which provides standards for interoperability among all health care stakeholders. The Clinical Document Architecture (HL7 CDA, 2010) is a XML-based HL7 standard for the representation and machine processing of

clinical documents for the purpose of exchange. CEN 13606 (CEN 13606, 2010) is an electronic health record (EHR) extract standard that is developed by the European Committee for Standardization (CEN) and uses the openEHR archetype methodology (OpenEHR, 2010). An approach to encoding doctors notes using SNOMED defined codes is described in (Patrick et al, 2007).

2.2. Data Integration

Data integration is an important process for public health surveillance. Privacy and security are sensitive issues when conducting disease surveillance. This section describes current theories and practices in data integration methodologies, architecture, and patterns, as well as related security and privacy mechanisms.

2.2.1 Cross-Industry Standard Process for Data Mining

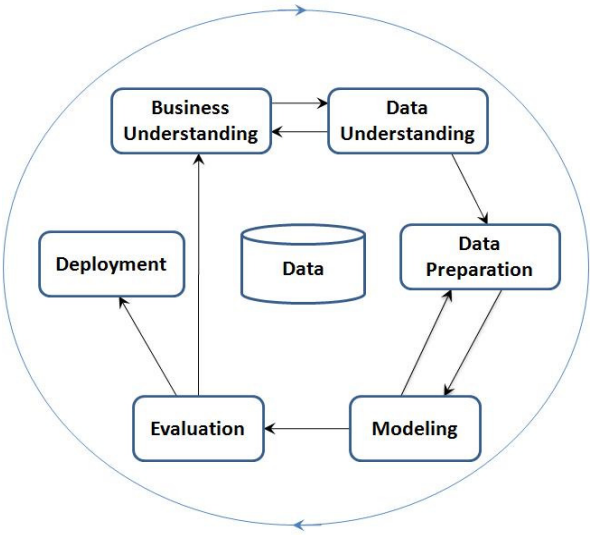

Figure 2-1 CRISP-DM (CRISP-DM, 2010)

Public health surveillance is usually conducted through a knowledge discovery or data mining process. Privacy-preserving data integration must be understood within the context of an overall methodology or process for data mining within which data

integration occurs. The European Union defined a standard methodology or process for tool-neutral data mining (Shearer, 2000; CRISP-DM, 2009), the Cross-Industry Standard Process for Data Mining (CRISP-DM). It organizes the data mining process into six phases when conducting a data mining project: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Data integration is planned during the data understanding, data preparation and modeling phases, and is executed on a continuous basis during the deployment phase. However, ethical and regulatory concerns relevant to data integration are also considered during the business understanding and evaluation phases.

However, the focus on CRISP-DM is largely on data mining projects in which data is collected within a single organization (possibly from many different sources). It does not address privacy or B2B issues which are important for public health surveillance.

2.2.2 Data Partition Model and Data Warehouse

Health care data is widely distributed across many different organizations. When attempting to integrate data from distributed data sources, there can be two different structured data partitioning models: vertically partitioned data and horizontally partitioned data. For horizontally partitioned data, the data subjects (rows) are partitioned among distributed databases that have the same sets of attributes for all of its subjects. For vertically partitioned data, the attributes (columns) are partitioned among distributed databases that contain only partial attributes of a subject. Both structured partitioning models engender significant metadata issues when integrating data. The horizontally

partitioned databases need to ensure that their data subjects are not duplicated. The vertically partitioned databases have difficulties in linking records across the databases.

A data warehouse (Kimball et al,1998; Inmon, 2005) consolidates all specified data into a centralized repository, often with a generalized, global schema. They are reliable and generally provide excellent response time to user queries. The traditional data warehouse is useful inside a single organization. There are two major design methodologies of data warehousing. The top down approach (Inmon, 2005) insists that data is organized into subject oriented, integrated, non volatile and time variant structures. The data marts are treated as sub sets of the data warehouse. Each data mart is built and optimized for particular analysis needs. The bottom-up approach (Kimball et al, 1998) designs the data warehouse with the data marts that are connected to it using a bus structure that contained all the common elements used by data marts such as conformed dimensions and measures.

A federated approach can be adopted in B2B health care networks. A federated data warehouse is one in which separate, distributed, heterogeneous data warehouses functionally act as a single one. It can help rapidly integrate data in real-time from disparate data sources. But it can be problematic when integrating data across organizations. A conceptual model of a federated system is introduced in (Stolba et al, 2006). The three-phase consolidation process (depersonalization, pseudonymization and federation) is used to ensure that personal identities are made secret before sending data to be federated. It uses a trusted third-party organization for pseudonymization and

linking records. However it will fail if the third party is adversarial or the result set has a high re-identification risk.

2.2.3 Data Mashups and Software as a Service

Health 2.0 refers to the transformation of the health care systems based on leveraging emerging Web 2.0 technologies, principles and practices (O'Reilly, 2004) for building applications and services based on the Internet. Data mashups are Web 2.0 technology that is relevant to data integration. They are based on integrating data from disparate data sources, services or data feeds published by a Web site. Often mashups are created in an ad hoc fashion. But privacy, authentication, compliance and other constraints must be taken into consideration for an enterprise mashup (Hanson, 2009). Software as a Service (SaaS) is a business model to deliver applications or software to the end user as a service. SaaS enables availability of B2B applications anywhere to end users with Internet connections as well as providing real-time support. However there are also risks in using mashups and SaaS in terms of their continued support, reliability, security, and scalability. In particular, a big issue is how to securely link and safeguard records across multiple sources in such a way that privacy and security requirements are addressed.

2.2.4 Data Integration Patterns

There are several examples in the literature that use patterns as the language to describe the implementations and the underlying design decisions of data integration. (Schwinn et al 2005) identifies five different data integration patterns: two redundancy-free solutions, and three redundancy-based solutions. A collection of security patterns is

summarized in (Yoshioka et al 2008, Kienzle et al, 2010) to provide solutions to recurring security problems. However, there is little discussion of privacy preserving data integration patterns that would be useful for public health surveillance.

2.3. Security and Privacy Mechanisms

This section introduces identity management and cryptographic systems that are useful for privacy protection in data integration.

2.3.1 Identity Management

Identity protection is one of the main goals of privacy protection in health care data integration. Anonymous and pseudonymous identities are two ways of protecting identity (Koch & Möslein, 2005). With anonymous identity, the identity is unknown and one cannot refer to an identity beyond the single session in which it is used. With pseudonymous identity, one can link events across sessions to an identity created for a specific organization, without knowing the actual identity. Anonymous and pseudonymous identity are key concepts in our proposed framework.

Federated identity management is a solution for managing identity across the organizations. OpenID (Recordon & Reed, 2006) is a free single-sign on solution originating with for internet users. Microsoft's Windows Live ID (Live, 2006) is also a single sign on service provided by Microsoft that allows users to sign on many websites only using a single account. A Circle of Trust is a key concept for protecting personal data that is shared between organizations over the Internet (Koch and Möslein, 2005). An architecture to support this has been developed by the Liberty Alliance project (Tourzan et al, 2006; Landau, 2003; Cahill et al, 2008). A Circle of Trust is a business to

business network in which an individual's identity and personal information is protected by a designated Identity Provider, while still allowing cooperating organizations within the Circle of Trust to access and share the individuals personal information over the Internet in a systematic manner. Sun has an implementation based on Liberty Alliance called OpenSSO (OpenSSO, 2010).

2.3.2 Secret Sharing and Homomorphic Cryptography

Privacy preserving data aggregation can be achieved by using a homomorphic encryption schema (Rivest et al, 1978), where specific arithmetic operations on the plaintext can be achieved by performing operations on the ciphertext. It ensures data confidentiality when computing encrypted data. There are several homomorphic cryptosystems such as RSA (Rivest et al, 1978), ElGamal (ElGamal, 1985), Goldwasser-Micali (Goldwasser & Micali, 1984), Benaloh (Benaloh, 1987) and Paillier (Paillier, 1999). They have been used in different application contexts: oblivious transfer mix-nets, watermarking, finger-printing protocols, electronic voting, auctions, lottery, and secure multiparty computation protocols.

Paillier Cryptosystem

Paillier encryption schema (Paillier, 1999) is frequently used in secure multiparty computation, privacy preserving data aggregation and data linking protocols. It is additively homomorphic and computationally efficient to decrypt. Paillier Cryptosystem has three steps:

1) Key generation
   - Pick two large primes p, q. And let $n = pq$
   - Create a public key (n, g) where g has order a multiple of n
   - Create a secret key λ where $\lambda = lcm(p-1, q-1)$.
     (lcm stands for Least Common Multiple)
2) Encryption
   - Let Z_n be the set of integers modulo n $\{0,1,\cdots,n-1\}$
   - To encrypt a message $m \in Z_n$, randomly choose $r \in Z_n^*$, which is the multiplicative group of invertible elements of Z_n
   - Compute an encryption $c = E(m) = g^m r^n \bmod n^2$
3) Decryption
   - Compute a decryption: $m = \dfrac{L(c^\lambda \bmod n^2)}{L(g^\lambda \bmod n^2)} \bmod n$,
     where $L(\mu) = \dfrac{\mu - 1}{n}$.

. It also has the following properties that are very useful in multiparty computation protocols: additive homomorphic, probabilistic encryption and semantic security. Given the public key and the encryption of m_1 and m_2, one can compute the encryption of $m_1 + m_2$:

$$E(m_1) \times E(m_2) = (g^{m_1} r_1^n \bmod n^2) \cdot (g^{m_2} r_2^n \bmod n^2) = E(m_1 + m_2)$$

$$D(E(m_1) \cdot E(m_2) \bmod n^2) = m_1 + m_2 \bmod n$$

Multiplication with a constant C can also be performed as follows:

$$E(m_1)^C \bmod n^2 = E(Cm_1)$$

$$D\left(E(m_1)^C \bmod n^2\right) = Cm_1 \bmod n$$

Secret Sharing and Threshold Paillier

Secret sharing is a technique by which a secret can be shared by multiple parties where no party knows the secret, but it is easy to reconstruct the secret by combining

each party's share. Shamir's secret sharing (Shamir, 1979) involves evaluations of a randomly generated polynomial. Threshold Paillier cryptosystem (Fouque et al, 2000) is based on secret sharing and is used when it is impossible to obtain all secret shares. The basic steps are:

1) **Initialization.** The KG creates the public, private and verification keys. The public key (n, g) and all the public verification keys, $v, v_i, i = 1,...,l,$, are publicized. A party X_i gets a secret share s_i of the private key s corresponding to the public key (n, g).
2) **Encryption.** Any party can run the encryption algorithm using (n, g).
3) **Decryption.** A party X_i decrypts the ciphertext c using his share of the secret key s_i to get the partial decryption c_i and forms a zero-knowledge proof of validity of the partial decryption. If a set of t or more parties have valid proofs, they can recover the plaintext.

Trust and efficiency are two main problems when a protocol uses a pure secret sharing technique such as Shamir's secret sharing. Other security mechanisms such as digital signature and homomorphic encryption need to be combined with it to achieve complex security requirements to make sure that the protocols are trustworthy and resist collusion among stakeholders.

2.3.3 Digital Signature and One Way Accumulator

A digital signature (ElGamal, 1985; Rivest et al, 1978) is an electronic signature that uses a mathematical scheme to authenticate the identity of a digital message or document. It requires the use of a Public Key Infrastructure (Diffie et al, 1996) technology and provides a level of trustworthiness and accountability that aids multi-party communication and computation in data integration. A digital signature scheme typically consists of three algorithms: a key generation algorithm; a signing algorithm;

and a signature verifying algorithm. The use of asymmetric digital signatures assures three functional properties: authenticity of the source of a message; integrity of a message and non-repudiation of a message.

One-Way Accumulator

Benaloh proposed a decentralized alternative to digital signatures (Benaloh et al, 1994), where a one-way hash function satisfying a quasi-commutative property that is used as an accumulator. The desired property is obtained by considering functions h: $X \times Y \rightarrow X$ and for all $x \epsilon X$ and for all $y_1, y_2 \in Y$, $h(h(x, y_1), y_2) = h(h(x, y_2), y_1)$. This one-way accumulator can be used for membership testing, pseudonym construction and authentication. It will enhance trustworthiness when it is used in a multi-party computation protocol.

2.3.4 Access Control

Access control is relevant to data integration since organization permissions are required to coordinate access for defining, building and accessing datasets. Role-based access control (RBAC) (Sandhu et al, 1996), team-based access control (TBAC) (Thomas, 1997), and policy-based control (Rouault & Clercq, 2004) are three common approaches to access control. The RBAC framework is based on users, roles and access permissions. Access permissions are reassigned easily from one role to another without modifying the underlying access structure. A "team" in the TBAC model provides fine-grained control over permission activation to individual users and objects. The policy-based control of access is based on three main components: The policy administration point, the policy decision point (PDP) and the policy enforcement point (PEP).

2.4. Privacy Preserving Record Linkage

Record linkage is a key component of data integration. It is a process of determining whether two records in different sources refer to the same individual. Privacy preserving record linkage (PPRL) techniques are often used for health record integration and aggregation.

2.4.1 Secure Multi-party Computation Techniques

Secure multi-party computation techniques (Fouque et al, 2000, Pinkas, 2002) based on cryptographic protocols provide provably secure solutions to privacy preserving data analytics. As introduced in Yao's Millionaire problem (Yao, 1982), it enables the computation of some global data characteristics among different data sources without revealing individual private data. Secure multi-party computation can be used in privacy-preserving protocols. Some protocols use third parties to perform linking or computation. These can be classified as trusted third party and semi-trusted third party protocols (Franklin et al 2010). Some protocols make no use of a third party.

- **Trusted Third Party** is fully trusted by two or more parties; and facilitates interactions among them. Using a Trusted Third Party is only secure if the Third Party is completely trusted. If the record linking task is performed by a Trusted Third Party, then the Third Party is trusted to see the intermediate results of the record-linking.
- **Semi-trusted Third Party** is a third party that follows the protocol as it is supposed to. However, unlike a Trusted Third party, they will store and use received intermediate values to infer private information if it is possible; for

example re-identification. Protocols which use a semi-trusted Party instead of a Trusted Third Party provide stronger security because they do not require the Third Party to be completely trusted, but rather ensure that private information cannot be inferred by them. Encryption or other techniques prevent the Third Party from seeing the actual intermediate values. The Semi-Trusted Third Party is only trusted not to take advantage of its participation in the protocol to maliciously try to hack or break the protocol.

Approaches using a Trusted Third Party

Current common approaches for record linking use one way hash function to encrypt identifiers and send them to a trusted third party for linking. There are several variations. (Galindo, 2010) built a pseudonymized data sharing system where electronic health records can be linked. The cryptographic implementation of the pseudonymization system satisfies ISO/TS 25237:2008 (Pseudonymization, 2008). ISO/TS 25237:2008 contains principles and requirements for using pseudonymization services to protect personal health information. ISO/TS 25237:2008 is applicable to organizations who make a claim of trustworthiness for providing pseudonymization services.

Federation and pseudonymous linking have been used to provide a unified view of the health care of population (Ainsworth et al, 2009). Other approaches leverage mediator-based architectures. (Boyens et al, 2004) focuses on the problem of interval inference while (Schadow et al, 2002) proposes that a particular set of real patient identifiers used for the joining be one-way-encrypted by a keyed hash-function.

Approaches using a Semi-trusted Third Party

It is theoretically possible to use techniques from cryptography (Goldreich et al, 2000) under a semi-trusted model. In (Karakasidis et al, 2009), a deterministic record linkage method is used that combines social security number, first name, birth month and gender as patient identifiers. After phonetic encoding of real and fake patient identifiers, the data sources send them to a semi-trusted third party that works as a classifier to perform the join operation and decide upon the matching status of the record pairs using binary field comparison. Sources contact each other directly asking for rows in which the fields have the resulting identifiers.

(Kantarcioglu et al, 2009) introduces two protocols to create an anonymous integrated encrypted database at a semi-trusted data storage site. The Secure-Equijoin protocol uses a cryptographic system to join the encrypted records on encrypted identifying attributes. k-Equijoin protocol combines k-anonymity and Secure-Equijoin to improve efficiency. Patient records are shared and linked.

A hybrid approach to private record linkage is proposed in (Inan et al, 2008). It combines sanitization techniques (k-anonymization) and cryptographic techniques (secure multi-party computation protocols). The methods assume three participants: two data holders with the datasets to be linked, and the querying party who provides the classifier that determines matching record pairs. This approach enables users to trade off between privacy, accuracy and cost.

Approaches without a Third Party

It is desirable that linkage protocols do not involve the release of any private information, even to a trusted or semi-trusted party. (Agrawal et al, 2003) restricts their protocol to only two parties. The database queries considered are: intersection, equijoin and intersection size. But the proposed protocols all involve the sender and receiver obtaining additional information about each other's databases and the answer to the query.

Record linkage already plays a large role as a building block for privacy preserving statistical analysis. (Karr et al, 2009) propose a protocol for conducting secure regressions and similar analyses on vertically partitioned data – databases with identical records but disjoint sets of attributes. The data owners strictly adhere to an established protocol designed to preserve privacy. No third parties are involved.

2.4.2 Inference Control Methods

Inference control or statistical disclosure control protects the released data or the result of a computation so that they cannot be linked to specific individuals or entities. There are at least three inference control methods (Domingo-Ferrer, 2007) for privacy preserving protocols. The first method uses aggregates; the second method involves dynamic database techniques including query auditing and restriction (Gopal et al 2002, Nabar et al 2006), interval answers (Gopal et al 1998, Garfinkel et al 2006), and perturbation of answers (Duncan et al 2007, Herranz et al 2010); the third method involves micro data protection techniques including perturbative masking and non-perturbative masking.

There are tradeoffs between information loss and disclosure risk. The following methods help to select proper parameters for the used method and achieve sufficient protection at minimum information loss: score construction (Domingo-Ferrer et al, 2001), R-U maps (Duncan, 2001), and k-anonymity (Sweeney, 2002-1, Sweeney, 2002-2).

2.4.3 Identity Matching Strategies

Privacy preserving record linkage involves matching identities from different data sources to support linking and aggregating while at the same time protecting or hiding identity to ensure compliance with privacy laws.

- **Deterministic matching** achieves a match only if the fields being compared are identical.
- **Probabilistic matching** uses a greater number of matching variables to provide a maximum likelihood estimate among potential matches.

To securely link identities, some approaches directly encrypt the identifiers and compare the ciphertext for equality based on equivalence testing (Agrawal et al 2003; Agrawal et al 2004, Berman et al 2004, Eycken 2000, Grannis, 2002, Karakasidis, 2009). Some approaches leverage encryption and similarity techniques to achieve probabilistic privacy preserving linking. Trigrams (Hylton et al, 1996) combine encryption and the 3-gram method. When comparing two strings, the difference of the number of times a trigram appears in each string is calculated and put into a vector. The similarity score is assigned based on this difference vector. In (Schnell et al, 2009), each string is encoded by hashing the bigrams associated with the string into a Bloom filter using q different hash function. To determine the approximated similarity of two strings, the corresponding

Bloom filters are compared using a set-based similarity measure, such as Dice coefficient, which indicates the similarity for the two Bloom filters representing the encoded strings.

In (Swire, 2009), IBM DB2 Anonymous Resolution applies a one-way hash function to transform selected multiple fields into cryptographic values, which are sent to a trusted third party Resolver to match records. Fellegi and Sunter (Fellegi, 1969) introduced a formal mathematical model for binary field comparison and probabilistic linkage. In this model a field comparison vector contains the similarity scores for each field in each record pair, and represents the conditional probability of field agreement given the match status of the record pair. The linkage score for each record pair is calculated upon the vectors. In (Durham et al, 2010) three record linkage approaches are compared: binary field comparison and deterministic linkage, binary field comparison and probabilistic linkage, and approximate field comparison and probabilistic linkage. The results showed that approximate field comparison and probabilistic linkage has greater precision than others.

Deterministic matching can provide precise and correct matching by using common identifiers. But in some situations, there is no common identifier or identifying information is inconsistent. A probabilistic matching needs to be used. However, for privacy preserving data integration, probabilistic matching approaches have some limitations. In practice, record linking applications generally rely on people to manually do final decisions for those probabilistically matched records. This cannot be done for privacy preserving record linkage since the clear records are not allowed to be reviewed

by people except the data custodians. Most approaches anonymize the personal identifier using a one-way hash function before sending out. Some use a unique identifier; some use a combination of identity info such as name, age, address or others. This approach cannot resist dictionary attack. Some approaches add a random string known as “salt” to the identifiers to try to prevent dictionary attacks. In order to achieve correct matches, different data sources use same salts for a record linkage task. When this happens, it is possible that the same individual will have the same hash value, which is not secure. Most privacy preserving record linkage protocols only consider privacy protection during the linking process, without taking into account re-identification after linking.

Chapter 3. Problem Definition

This chapter describes three health surveillance scenarios to illustrate the problems we want to addresses; and then uses a gap analysis of the relevant literature based on the scenarios to identify evaluation criteria that should be addressed in any proposed solution.

3.1. Data Integration Scenarios

In developing this thesis, we identified three different types of data integration for public health surveillance. We illustrate the three types with scenarios in this section:

- Data aggregation for syndromic surveillance
- Pseudonymous record linking for adverse event tracking
- Anonymous data linking and aggregation for population based surveillance

These scenarios are also useful for our gap analysis of the literature and for understanding the key criteria that should be addressed to bridge that gap.

3.1.1 Data Aggregation for Syndromic Surveillance

It is important to facilitate effective and secure integration of appropriate information from hospitals, laboratory and health care practices to detect new disease outbreaks rapidly. The Internet and secure B2B networks offer the possibility of providing near real-time data integration. For many disease surveillance systems, counts of patients presenting with the particular syndrome under surveillance are required to detect meaningful trends. Individual patient level data is generally not needed to detect

epidemics, but some demographic information is helpful to detect and localize epidemics within geographically or demographically defined sub-populations. Nonetheless, although individual patient level data is not used, care must be taken to ensure that individual patients cannot be identified from the aggregates (especially where totals are small).

Figure 3-1 shows a system for public health agencies to detect epidemics and identify interesting trends. Each provider determines its case and patient counts and then submits them to public health authorities. Aggregate totals are collected at local or municipal health authorities, state or province authorities, and finally at national or international authorities. Once an epidemic is detected, public health agencies often require the ability to identify interesting cases for follow-up.

In this scenario, concerns about privacy and confidentiality often limit the willingness of data custodians to share data. These concerns extend beyond preserving patient privacy to include wishing to protect the identity of the data providers who supply the data. A doctor office, clinic or hospital does not necessarily want to be known as affected by the outbreak. Although aggregate data can be shared if the data is de-identified, the reporting organizations are often unsure of the risk exposure if they disclose patient data on an on-going basis and may be concerned about how data about their organization may reflect on them.

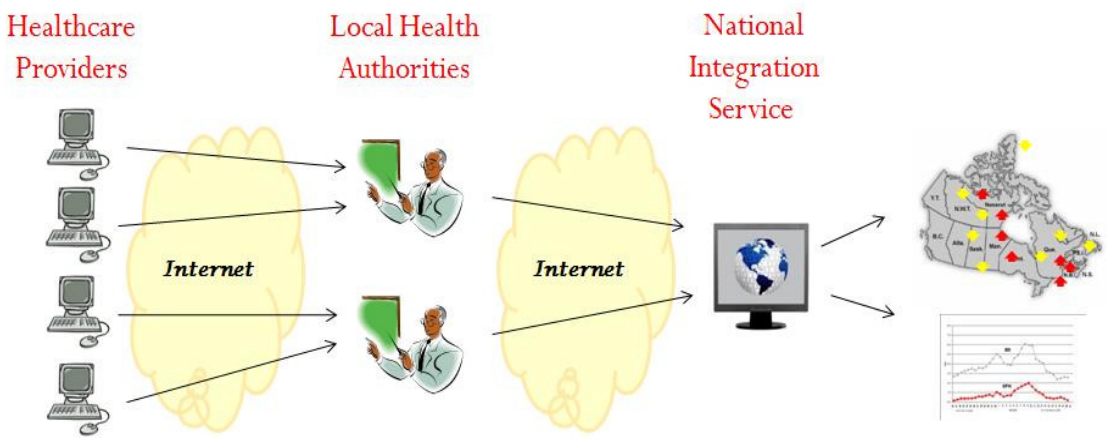

Figure 3-1 Syndromic Surveillance

3.1.2 Record Linking for Adverse Event Tracking

In this scenario, medical researchers would like to do an analysis to discover if there are any patterns that might correlate emergency room visits to symptoms or prescriptions through a knowledge discovery process upon a consolidated data set. In the "Consolidated Data Set" in Figure 3-2, each column of data comes from a different service. For a single patient, the Clinic only knows the patient's symptoms; the Pharmacy only knows the patient's prescription drug information, and the Hospital only knows the patient's emergency room event. A mechanism is needed to link these three databases to create a consolidated and composite view of a single patient. There are a number of issues. First of all, it is not allowed to create such as dataset without appropriate privacy safeguards. Secondly, there is no standard identifier among these databases. Different organizations build and use their databases separately. They may use different types of identity information and the information may be incomplete.

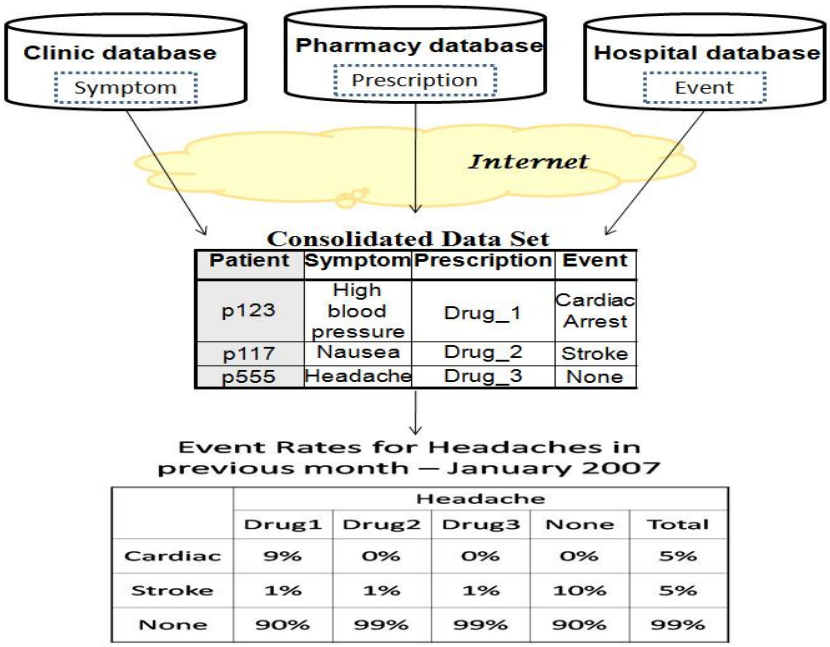

Figure 3-2 Adverse events tracking

3.1.3 Anonymous Data Linking and Aggregation for Population Health Surveillance

HPV, human papillomavirus, is one of the most common sexually transmitted viral infections in the world (Clifford et al 2005). A comprehensive HPV surveillance system would like to track the relationship of HPV to cancer rates as well as the effectiveness of immunization and public awareness campaigns. Four separate data sources provide the information: a PAP test registry for early detection of cervical cancer, a HPV typing lab database for HPV infections, a sexual health survey to gauge the effectiveness of public awareness campaigns, and a vaccine registry which records who has received the vaccine for HPV. In many jurisdictions, the privacy of these four data sources is strictly protected, and no record linking based on patient identity (not even pseudonyms) is allowed. Data needs to be integrated and aggregated from these sources anonymously without sharing patient identity in order to investigate trends and

relationships among different variables based on statistical computations. The sample reports are shown in Figure 3-3. Table 1 shows the relationship between HPV types and cytological results which need to link two data sources. Table 2 is a report linking HPV data registry and a survey database that contains demographic information. Table 3 is a report of statistics for analyzing relevant factors.

Table 1. HPV prevalence by type and cytological outcome

HPV	Missing		Normal		Unsatisfactory		ASC-US		LSIL		ASC-H		HSIL		Total	
types	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
negative	10	71.4	440	81.3	17	85.0	12	70.6	7	38.9	3	75.0	2	33.3	491	79.2
any	4	28.6	101	18.7	3	15.0	5	29.4	11	61.1	1	25.0	4	66.7	129	20.8
6 or 11	0	0.0	15	3.4	2	11.8	0	0.0	1	14.3	1	33.3	1	50.0	20	4.1
16 or 18	1	7.1	26	4.8	1	5.0	0	0.0	3	16.7	1	25.0	2	33.3	34	5.5
low-risk	0	0.0	41	7.6	2	10.0	2	11.8	3	16.7	1	25.0	2	33.3	51	8.2
high-risk	2	14.3	61	11.3	2	10.0	2	11.8	8	44.4	1	25.0	4	66.7	80	12.9

Table 2: Survey results

Variables	Categories	Age < 30		Age ≥ 30	
		HPV - N =125	HPV + N = 86	HPV - N = 433	HPV + N = 70
Ethnic identity	Aboriginal	26.2	36.1	16.6	32.7
	white	49.4	41.6	67.4	42.8
	other	12.6	4.7	5.8	7.8

Table 3: Factors associated with the risk of having a HPV positive Pap test

Variables	Category	Odd ratio	95% CI	P-value
Age	(continuous)	0.96	0.95 - 0.99	0.002
Ethnicity	other	1	--	
	Aboriginal	3.17	1.11 - 3.12	0.005
	two or more	4.99	4.56 - 12.23	<0.0001

Figure 3-3 Sample HPV Reports (Note: data is fictional)

3.2. Summary of Gap Analysis

This section presents a summary of a gap analysis that was done of data integration requirements for health surveillance illustrated in our scenarios against a survey of the mechanisms and approaches described in the literature survey in Chapter 2. As discussed in Section 2, federated identity, access control, encryption and intelligent monitoring are critical to protect privacy, but they do not directly provide data integration. Web services,

data warehouse, and data mashups are useful for B2B data integration but privacy protection must be integrated into them. The following highlights the gaps between the requirements and existing solutions from different aspects.

3.2.1 Gap in Data Integration

We have investigated and analyzed the approaches related to data integration in Section 2.2. Table 3-1 is a summary of gaps in existing data integration solutions.

Table 3-1 Summary of gaps in existing data integration approaches

Approach	Gap
Health research approval process	No support for near real time data integration.
CRISP-DM	No support for B2B data integration; Does not address privacy and security issues.
Traditional data warehouse	No support for B2B data integration; Does not address privacy issues.
Data mashup and SaaS	Does not address privacy and security issues.
Security design patterns	No support for B2B data integration; No support for near real time data integration.

3.2.2 Gap in Security and Privacy Mechanisms

We have investigated and analyzed the approaches related to security and privacy mechanisms in Section 2.3. Privacy cannot be sufficiently protected by privacy legislation or privacy codes of conduct alone. Security mechanisms are useful for privacy protection, but current existing solutions do not support the concept of data integration. Table 3-2 is a summary of gaps in existing security solutions.

Table 3-2 Summary of gaps in existing privacy solutions

Approach	Gaps
Circle of Trust	Difficult to set up; No support for bulk data integration. Does not completely ensure that identity info is kept secret.
Pure secret sharing techniques	Does not address data integration; Does not address trust and data authentication.
Homomorphic cryptography	Does not address data integration;
Access control	Does not address data integration;

3.2.3 Gap in Privacy Preserving Data Linkage

Data linkage is a key concept in data integration. We have investigated and analyzed the approaches related to privacy preserving data linkage in Section 2.4. Table 3-3 is a summary of gaps in existing privacy preserving data linkage solutions.

Table 3-3 Summary of gaps in existing privacy preserving data linkage solutions

Approach	Gaps
Using a trusted third party	Does not completely ensure that identity info is kept secret since it requires full trust of a third party.
Using semi-trusted third party	Does not address trust and data authentication.
Using no third party	Requires providers to completely trust each other.
Deterministic matching	Does not support inconsistent identifiers.
Probabilistic matching	Does not support precise integration and has limitations on correcting errors.

Linking a hashed identifier	Does not resist dictionary or collusion attacks.
Most linking protocols	Do not consider re-identification risk after linking

3.3. Evaluation Criteria

Privacy protection and data integration are vital but often conflicting aspects of public health surveillance. Privacy requirements should be technically enforced and designed for data integration in a manner that still supports full aggregation and linking so that data can be integrated across a health care network. Table 3-4 lists a set of evaluation criteria that can be used to analyze any framework or proposed solution for addressing and managing this issue. The table was created based on

- Our gap analysis using our scenarios and literature survey described in chapter 2;
- Two case studies done in collaboration with public health organizations (which involved the three types of data integration illustrated by our scenarios in this chapter);
- Discussions with experts from CHEO research institute, the University of Texas at Dallas, International Organization for Surveillance and Public Health Agency of Canada;

We do not claim that this is a complete set of the criteria that a framework for privacy preserving data integration must meet, but it does reflect the essential criteria in terms of the current gaps we have identified.

Table 3-4Essential criteria for Data Integration Framework in public health surveillance

Aspect	Criteria
1. Data integration	a. Support distributed data sources cross organizations. b. Support near real time data integration. c. Enable data publishing and data reporting.
2. Privacy of patient	a. Ensure that the patient identity is kept secret. b. Ensure that integrated data cannot be re-identified. c. Ensure that patient consents and/or organizational agreement are in place.
3. Identity linking	a. Ensure linking identity without revealing identity info.
4. Privacy of data provider	a. Ensure that the data provider identity is kept secret. b. Ensure that integrated data cannot be re-identified.
5. Data Protection	a. Ensure integrity of data. b. Authenticate sensitive data sources. c. Prevent adversary attacks such as organization collusion attack. d. Control Access to sensitive data.

3.3.1 Data Integration

The data integration framework should create data sets assembled from data sources across a federated health care network. For example, a common data view is required to consolidate and integrate data from multiple data sources. A metadata model or a data dictionary is also needed to define attributes and map them to data sources. As well, the

framework should support near real time and robust integration to provide timely, precise, and integrated data reporting for further analysis. This is important because today there is a lot of data either not available or available too late, or not delivered to the right people.

3.3.2 Privacy of Patient

The data integration framework should protect patient identity and safeguard privacy at different levels. First, data collection should conform to privacy legislation. In particular, patient consents and/or organizational agreements should be obtained before their data is collected. Second, the patient identity should be kept secret from organizations in a B2B network and consumers of the data integration service. Technical mechanisms can be used to mask or hide identity and sensitive data through depersonalization, anonymization and psedonymization. Thirdly, the integrated data should resist re-identification attack and it should not be possible to re-identify a patient.

3.3.3 Identity Linking

In some circumstances, the data integration framework should allow identities to be linked for creating consolidated data sets or obtaining specific reports. They can be linked deterministically or probabilistically, directly or by calculation. In all cases, patient identity should be protected when linking.

3.3.4 Privacy of Data Provider

The data integration framework should protect the privacy of data providers. In the process of integrating data, the data provider identity should be kept secret from organizations in a B2B network and consumers of reports from the data integration

service. After data integration, the integrated data should resist re-identification attack for data providers.

3.3.5 Data Protection

The data integration framework should provide privacy enhancing security mechanisms to ensure data security. A mechanism should be established to authenticate sensitive data sources to ensure that data is sent to and comes from valid entities. Some techniques should also be used to ensure integrity of data so that data is not visible to unauthorized users and protected against tampering. Access control mechanisms can be put in place to ensure sensitive data is accessed only by authorized users. The framework should prevent adversary attacks such as a network attack or an organization collusion attack.

Chapter 4. New Data Integration Protocols

Based on the gap analysis in Section 3.4, we develop three new data integration protocols using algorithms that had not been applied to data integration for public health surveillance. The protocols are described in the following sections.

4.1. A Protocol for Provider Anonymized Aggregation

This section describes a protocol for protecting the identity of providers in syndromic surveillance as described in Section 3.1.1. Much of what is described here was first introduced in (El Emam, Hu & et al, 2010, under second review at JAMIA). This protocol is the "first" protocol in the literature to protect privacy of the providers.

4.1.1 Requirements

As described in Section 3.1.1, the health researchers wish to track trends in symptoms potentially related to some disease across Canada and eventually the world in near-real time, in order to provide dashboard and statistical views of how the disease is spreading. An integration service is available over the Internet that collects, integrates and reports the counts of patients from health care providers such as doctor's clinics and emergency rooms. The health care providers send their counts of patients and cases to the local health authorities (usually municipal); then the local health authorities summarize their counts and send to the integration service. However, participation in this service is low, because health care providers have privacy, confidentiality and trust concerns about submitting such data electronically.

Imagine that there are M different organizations (providers) across Canada who are willing to share their case and patient counts for disease surveillance if patients and providers will not be identified when the surveillance results are public. Each provider determines its case and patient counts for each stratum. Each stratum is defined by a syndrome and an age group. These providers are clustered into groups of at least 5 sites based on their geographical locations. The counts within one group will be summed up using a secure computation of counts and rates protocol. Finally the aggregated counts and corresponding rates are displayed on a map.

To address the privacy concern of the providers in the above situation, the protocol should meet the following requirements:

1. The protocol should be able to aggregate horizontally partitioned data across organizations.
2. The protocol should protect the identity of patients.
3. The protocol should protect the identities of providers.
4. The protocol should handle re-identification risk for the location based reporting.
5. The protocol should prevent any single third party or adversarial party to know with certainty the true data from a data provider.
6. The protocol should prevent any single third party or adversarial party to know with certainty the true data for a patient.
7. The protocol should be trustworthy and ensure that the data comes from valid providers and not be altered.
8. The protocol should be robust to tolerate some technology failures.

4.1.2 Protocol

We develop a k-key holder protocol that leverages multi-party secure computation technique, threshold Paillier homomorphic cryptography and digital signature mechanism. Paillier cryptography maintains confidentiality of data and enables aggregation operation performed on encrypted value so as to protect privacy of provider; multi-party secure computation prevents any single third party to know with certainty the true data from a data provider; digital signature schema enhance trustworthy among all stakeholders. Figure 4.1 illustrates our approach where there are five main entities:

- **Providers**: are health care providers who calculate the counts of patients and cases encrypt them, and send them to the aggregators. The minimum number of Providers is five so as to prevent the risk of re-identification.
- **Aggregators**: are semi-trusted third parties who sum the encrypted counts by grouping the providers based on their regions. The minimum number of Aggregators is one; but we use two (or more) Aggregators to ensure the system is robust and fault tolerant.
- **Key Holders**: are semi-trusted third parties who decrypt the sums from the aggregators. The minimum number of Key Holders is two; but we use three (or more) to prevent collusion attack among Key Holders.
- **Data Integration Service**: is a semi-trusted third party who combines the partial decryption from the key holders.
- **Key Generator**: is a trusted third party who generates the public and private keys used in the protocol.

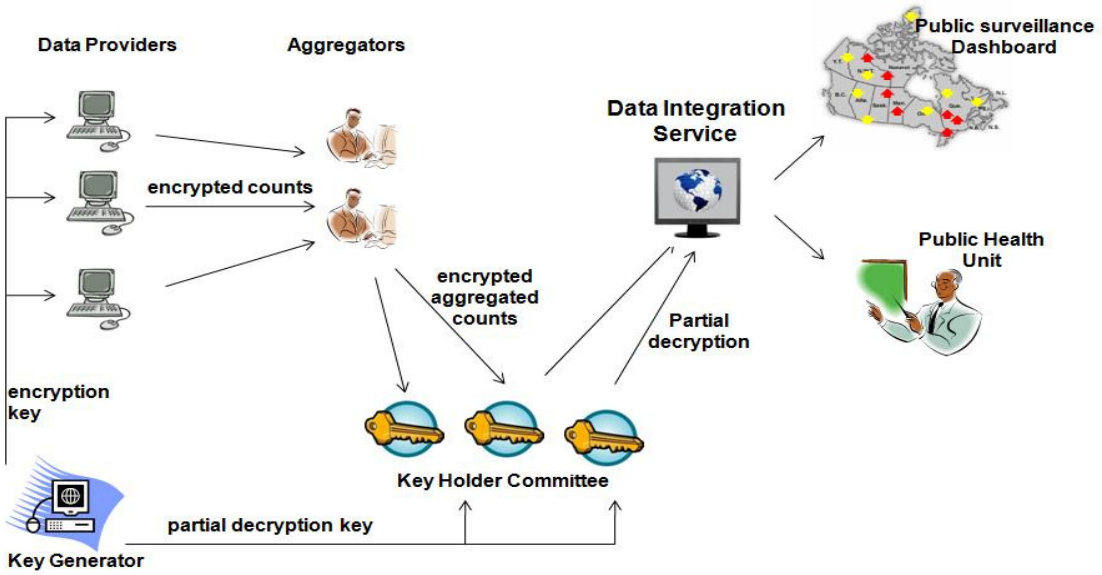

Figure 4-1 Overview of privacy preserving integration protocol

We use two aggregators which are fully redundant so that the protocol will work well even if one of the aggregators fails. We use three key holders so that the protocol still obtains the correct results even if one of the key holders fails. This protocol protects against the possibility of some parties colluding with each other. Especially it prevents any single adversarial party from knowing with certainty the true values for any data provider. It is also resistant to collusion between any aggregator and any one key holder. The complete protocol contains five main steps which are illustrated as follows. Formally we assume there are P strata, M providers, N aggregators, T key holders and R regions.

1. Setup

1) Each provider and each aggregator registers with the key generator (KG) for joining the surveillance system.
2) KG groups M providers into R groups, where each group has at least 5 providers. It also sends this information to each aggregator.
3) KG generates the following key pairs:

- Paillier public key PK, private keys SK_t and verification key VK_t, and send PK to each provider for encrypting the counts, and SK_t, VK_t to each key holder $t \in \{1,\ldots,T\}$ for partially decrypting the counts.
- Total M key pairs $<PK_i, SK_i>$ for each provider $i \in \{1,\ldots,M\}$ to sign the counts.
- Total N key pairs $<PK_j, SK_j>$ for each aggregator $j \in \{1,\ldots,N\}$ to sign the aggregated counts.

2. Each provider $i \in \{1,\ldots,M\}$ prepares and submits the counts.

1) Determine its case and patient counts C_{pi} for each stratum $p \in \{1,\ldots,P\}$. There are P different counts.
2) Encrypt each count using PK: $E_{pi} = E(C_{pi}, PK)$
3) Sign each encrypted count E_{pi} using SK_i: $SiE_{pi} = Sign(E_{pi}, SK_i)$
4) Send P encrypted counts E_{pi} and corresponding P signatures SiE_{pi} to each aggregator

3. Each aggregator $j \in \{1,\ldots,N\}$ verifies and sums up the encrypted counts.

1) Verify the signed encrypted counts from each provider $i \in \{1,\ldots,M\}$ using PK_i.

   If $Varify(E_{pi}, SiE_{pi}, PK_i) = Accept$, then the encrypted counts E_{pi} is valid.

2) Sum up the encrypted counts for each region $r \in \{1,\ldots,R\}$.

   $SUM_{pr} = \prod_i E_{pi}$ if provider i in Region r

3) Sign each aggregated encrypted count using SK_j: $SiSUM_{pr} = Sign(SUM_{pr}, SK_j)$

4) Send $R*P$ aggregated encrypted counts SUM_{pr} and corresponding $R*P$ signatures $SiSUM_{pr}$ to each key holder.

4. Each key holder $t \in \{1,\ldots,T\}$ verifies, decrypts the aggregated encrypted counts, and obtains the partially decryption.

1) Verify the signed aggregated counts from each aggregator $j \in \{1,\ldots,N\}$ using PKj

   If $Varify(SUM_{pr}, SiSUM_{pr}, PK_j) = Accept$, then SUM_{pr} is valid.

2) Decrypt each aggregated encrypted counts from each aggregator using SK_t.

$$DS_{prt} = D(SUM_{pr}, SK_t)$$

3) Generate proofs of validity of the partial encryption using VK_t.

$$VS_{prt} = \Pr oof(SUM_{pr}, DS_{pr}, VK_t)$$

4) Send DS_{prt} and VS_{prt} to the data integration service.

5. The data integration service combines partial encryption and obtains final counts.

1) Validate the proofs of correct decryption from each key holder $t \in \{1,\ldots,T\}$ using VK_t.

   If $Validate(VS_{prt}, DS_{prt}, VK_t) = True$, then DS_{prt} is valid.

2) Use (2, T) threshold Paillier algorithm and run combining algorithm to obtain the final counts: S_{pr} for each stratum $p \in \{1,\ldots,P\}$ and each region $r \in \{1,\ldots,R\}$

4.1.3 Security Analysis

The secure computation protocol adopts the idea of secret sharing, homogenous property of Paillier cryptosystem and digital signature. Table 4-1 is a summary of how the protocol meets the requirements identified in Section 4.1.1.

Table 4-1 Security analysis for secure computation of counts

#	Requirement Description	Protocol Description
1	Aggregate horizontally partitioned data	Each provider determines counts and the protocol obtains the total in the end.
2	Protect the identity of patients	No provider submits identity information about a patient
3	Protect the identities of providers	No provider submits identity information about a provider
4	Handle re-identification risk	The providers are grouped based on their regions where at least 5 providers are in a group.
5	Prevent any single third party or adversarial party to know with certainty the true data from a data provider	An aggregator only knows encrypted values from a particular provider. A key holder only knows partial decrypted counts across a group of providers. The data integration service only knows the total from all providers. And collusion attack is prevented by using multiple parties.
6	Prevent any single third party or adversarial party to know with certainty the true data for a patient	No provider submits individual patient level data
7	Be trustworthy and ensure that the data comes from valid providers	Leverages the digital signature schema
8	Be robust to tolerate some technology failures	Uses redundant aggregators and multiple key holders.

4.2. A Protocol for Master Patient Index Linking

This section describes a protocol for federated pseudonymous linking using a Master Patient Index. The initial idea for this protocol was first introduced in (Hu & Peyton, 2009).

4.2.1 Requirements

As described in Section 3.1.2, the researchers would like to analyze the relationship among the attributes which are vertically partitioned in different database across organizations. For example to track adverse events may require data from a clinic, local emergency room and a pharmacy. The databases from the three organizations may not be standardized on patient identifiers. Patient identity must be protected and privacy laws must be complied with throughout the process of matching identities and linking and integrating data sets. Furthermore, the analysis is continuous and, it would be good that the linking mechanism is reusable.

In short, the protocol should meet the following requirements to address the data integration and privacy concern in the above situation:

1. The protocol should enable integration of vertically partitioned data across organizations.
2. The protocol should protect the identity of patients.
3. The protocol should maintain consent control.
4. The protocol should handle identity inconsistency across organization.
5. The protocol should protect identity and health data during the integration process.
6. The protocol should maintain a reusable linking mechanism to improve efficiency.

4.2.2 Protocol

We develop a protocol that leverages a Master Patient Index to integrate data across organizations using federated pseudonyms. Figure 4-3 shows an overview of this protocol where there are three roles:

- **Providers**: are health care providers who provide health data in answer to queries but only using a pseudonym specific to them.
- **Identity Provider**: is a trusted third party who holds a Master Patient Index and performs pseudonym mapping and then send the destination pseudonyms to the data integration service. Figure 4-2 is an example of a Master Patient Index.

Master Patient Index

User_ID	Data Provider	Pseudonym
U1	Provider 1	U1_1
U1	Provider 2	U1_2
U1	Data Integration Service	U1_d
U2	Provider 1	U2_1
U2	Provider 2	U2_2
U2	Data Integration Service	U2_d

Figure 4-2 Example of the Master Patient Index

- **Data Integration Service**: is a trusted third party who sends request, receives and consolidates the results. A data set registry is required for creating a data dictionary for a federated query.

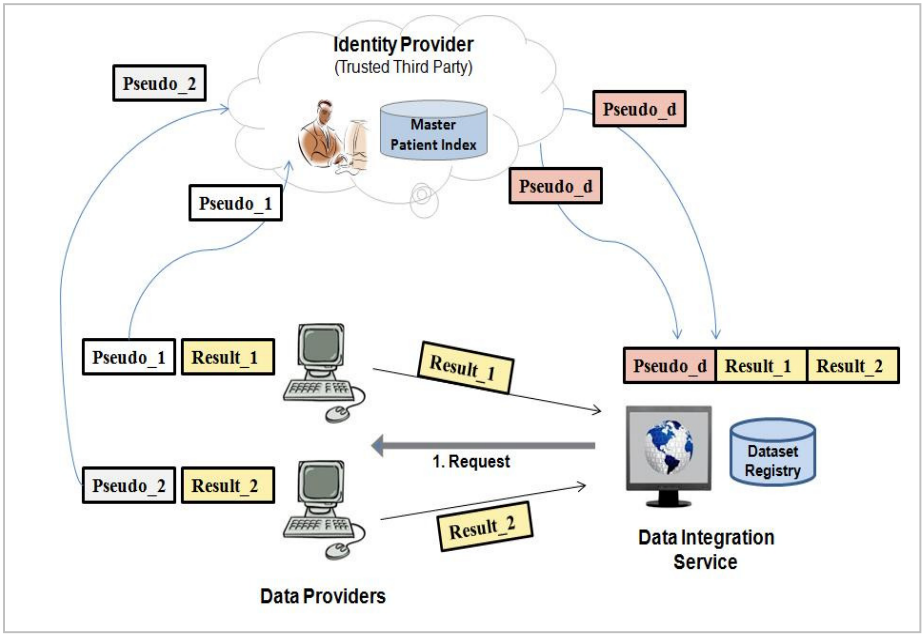

Figure 4-3 Overview of Master Patient Index linking Protocol

1. Setup

1) Participants register with Master Patient Index

   a. Each data provider registers with Master Patient Index (MPI) and receives some token to identify it to MPI.

   b. Each patient registers with MPI and receives a token to identify it to MPI. In the meanwhile, the consent for disclosing data can be obtained.

   c. Each provider registers patients with MPI providing it's token and the patient' token in order to receive a pseudonym for the patient specific to this provider.

   d. Optional: DB2 Anonymous Resolution (AR) (Swire 2009) can be used as a technique to populate the MPI. Figure 4-4 is a process of mapping identity using AR. If AR finds the coming identity already exists in the Master Patient Index,

the associated match in MPI is updated. Otherwise, a new pseudonym is issued to the coming identity.

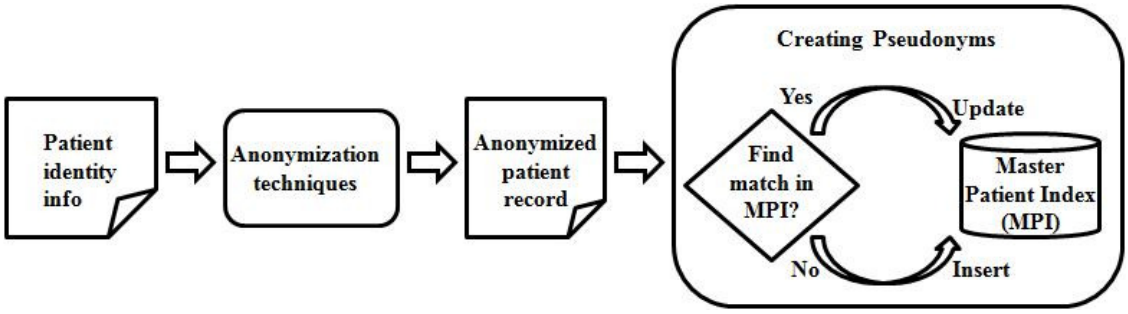

Figure 4-4 Identity mapping using AR

2) Each data provider registers with the data integration service that what the data is available and how to access it. This information will be stored in a dataset registry.

2. The data integration service sends requests to the data providers.

1) Receive a specific consolidated data set request.
2) Retrieve the sub-queries for from the dataset registry and ensure that all approvals have been obtained.
3) Send the sub-queries to related *M* data providers.

3. Each provider *i* $\in$ *{1,…,M}* executes the query and return results.

1) Process its corresponding sub-query and create a dataset indexed by its set of pseudonyms. The result includes two files. One only contains the index of pseudonyms: *Pseudo_i*. The other contains the attribute values that correspond to each pseudonym: *Result_i*
2) Send the pseudonym file *Pseudo_i* to the Identity Provider.
3) Send the attribute values *Result _i* to the data integration service.

4. Identity Provider converts pseudonyms and send to the data integration service.

1) Use the Master Patient Index to resolve the *Pseudo_i* and transform them into *Pseudo_d.*
2) Send *Pseudo_d* to the data integration service.

5. The data integration service consolidates all partial query results.

1) Join the pseudonym file *Pseudo_d* and all attribute values from all data providers to create the required consolidated dataset.

4.2.3 Security Analysis

The pseudonymous linking protocol uses pseudonyms to mask patient identity and leverage Master Patient Index to link the identities. Table 4-2 is a summary of how the protocol meets the requirements identified in Section 4.2.1.

Table 4-2 Security analysis for pseudonymous linking

#	Requirement Description	Protocol Description
1	Integrate vertically partitioned data cross organizations	Data registry maintains information of multiple sources and data is linked by Master Patient Index.
2	Protect the identities of patient	Identity of patient is masked using federated pseudonyms.
3	Maintain consent control	The providers are grouped based on their regions where at least 5 providers are in a group.
4	Handle identity inconsistency	Leverages AR probabilistic matching feature when a patient registers with MPI.
5	Protect identity and health data during the linking process	Pseudonymous linking using Master Patient Index and deterministic matching when consolidating data sets.

4.3. A Protocol for Secure Multi-party Computation Linking

This section describes a protocol for privacy preserving linking and secure computation across registries as described in Section 3.1.2. In particular, this protocol is useful when data providers are not willing or not authorized to share patient identity and the protocol illustrated in Section 4.2 is not allowed. Much of what is described was first introduced in (El Emam, Hu & et al, 2010, under review at JAMIA).

4.3.1 Requirements

As described in Section 3.1.3, the researchers would like to analyze the relationship among the attributes which are vertically partitioned in different disease registries across organizations. They want to generate the statistical reports based on the

data from multiple registries. However the sharing of identifying information is not possible because the registry owners are not willing or not authorized to share identifiable patient data. Therefore there has to be a mechanism for matching identities between different locations in order to link and aggregate data for a particular view or query, without sharing of any identity information. Moreover it would be better to resist malicious adversaries when the data integration service is deployed on the Internet. Further, one must ensure that combining data sets from different data sources into a single consolidated data set does not create data that may be potentially re-identified even when only summary data records are created.

To address the privacy concerns in the above situation, the protocol should meet the following requirements:

1) The protocol should be able to link vertically partitioned data from multiple registries and calculate the aggregates or statistics
2) The protocol should protect the identity of patients.
3) The protocol should handle re-identification risk for the population-based reporting.
4) The protocol should prevent for any single third party or adversarial party to know with certainty the true final value to be calculated.
5) The protocol should be trustworthy and ensure that the data comes from valid registries.

4.3.2 Protocol

We develop a secure multi-party computation protocol that leverages one-way accumulator mechanism (Benaloh et al, 1994) and Paillier homomorphic cryptography (Paillier, 1999). One-way accumulator can protect the identity of patient as well as link identities cross multiple registries. It is also used for membership testing. Paillier cryptography and multi-party computation enables to perform aggregation operation over encrypted values and prevents any single third party to know with certainty the true data for a patient. The protocol is illustrated in Figure 4.5 where there are three roles:

- **Registries**: are the data custodians that need to be securely linked. They are not allowed to share data on identity of patients.
- **Aggregators**: are semi-trusted third parties who securely link the registries and compute the query results and send to the data integration service. The minimum number of Aggregators is two so that any single Aggregator does not know with certainty the true data for a patient.
- **Data Integration Service**: is a semi-trusted third party who is on behalf of the end user sends request, receives the results and computer the final results.

We use two aggregators so that no single aggregator knows the real value of a count or a statistical data. In addition, one-way accumulator plays two important properties in our protocol: one is to securely compute and link the data from registries; the other one is to enhance the trustiness of the protocol by providing a mechanism of membership testing.

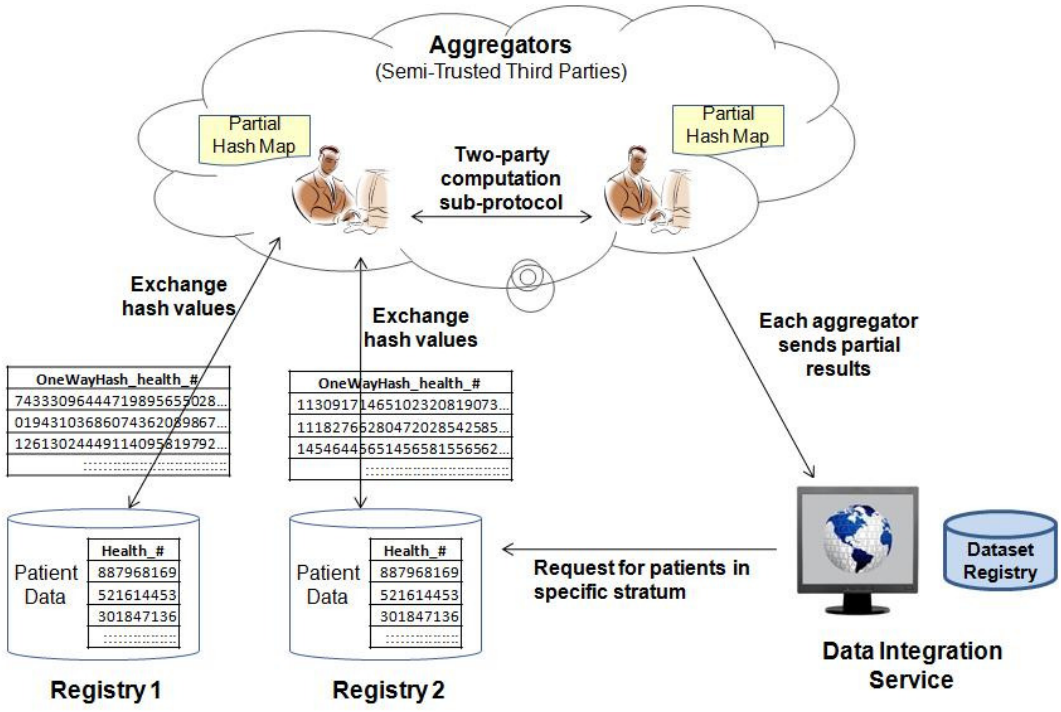

Figure 4-5 Overview of secure multi-party computation linking protocol

1. Setup

1) Each registry registers with the data integration service what the data is available and how to access it, as well as related access control policies. This information will be stored in a dataset registry. For example, Registry 1 can provide hashed patients with low-risk or high-risk HPV types. Registry 2 can provider hashed patients with normal or abnormal test results.

2) All registries agree upon a one-way accumulator *H* (Benaloh et al, 1994) that is used for performing hashing or rehashing the query results.

3) All registries agree upon a unique identifier for the patients. It can be the health insurance card number; or patient's name; or the combination of name, gender and address; or even an identifier generated by a tool that is able to handle identity inconsistent problems.

2. Data integration service sends requests to the registries.

1) Define contingency tables and corresponding queries.

Based on analysis requirements, the data integration service decides the queries to be requested. For example, a 2x2 contingency table is required to show the counts from Registry 1 and Registry 2. As shown in Figure 5-6, C_{11} is the count of patients who have low-risk HPV but the pap test results are normal; C_{22} is the count of patients who have high-risk HPV and the pap test results are abnormal.

	Pap Test Normal	**Pap Test Abnormal**
HPV Low-risk	C_{11}	C_{12}
HPV High-risk	C_{21}	C_{22}

Figure 4-6 Example of a contingency table

Four queries are defined for this table. Q_{1+} is for the low-risk HPV patients; Q_{2+} is for the high-risk HPV patients; Q_{+1} is for the patients with normal pap test results; Q_{+2} is for the patients with abnormal pap test results. Without loss of generality, we only illustrate how to obtain the result of C_{12}. It must be calculated based on the answers for Q_{1+} and Q_{+2}.

2) Check the dataset registry to see how to access the data from Registry 1 and Registry 2.
3) Generate a random number RQ_{12} for the cell C_{12}, which is used for membership testing.
4) Send RQ_{12} and Q_{1+} to Registry 1; send RQ_{12} and Q_{+2} to Registry 2, and refer these queries to C_{12}.

3. Each registry calculates the proof of membership and responds to the queries by sending hashed patients to aggregators.

1) Each registry generates a random number only known to itself, R_k, for each query Q_k.

2) Registry 1

   a. Calculates the proof of membership $H(R_{1+}, RQ_{12})$

   b. Responds with a hashed patients matching the query Q_{1+}. A hashed patient is notated as $H(R_{1+}, ID_i)$, where ID_i is the identifier for the patient from Registry 1

   c. Randomly selects an aggregator, for example Aggregator 1, and sends $H(R_{1+}, RQ_{12})$ and $H(R_{1+}, ID_i)$ to it. Registry 1 may send the hash value for the next patent to Aggregator 2

3) Registry 2

   a. Calculates the proof of membership: $H(R_{+2}, RQ_{12})$

   b. Responds with hashed patients matching the query Q_{+2}. A hashed patient is notated as $H(R_{+2}, ID_j)$, where ID_j is the identifier (such as the health insurance card number) for the patient from Registry 2.

   c. Randomly selects an aggregator, for example Aggregator 1, and sends $H(R_{+2}, RQ_{12})$ and $H(R_{+2}, ID_j)$ to it. Registry 2 may send the hash value for the next patent to Aggregator 2

4. Each aggregator forwards information among registries.

1) Aggregator 1

   a. Forwards the information to Registry 2 after it receives $H(R_{1+}, RQ_{12})$ and $H(R_{1+}, ID_i)$ from Registry 1, and tells Registry 2 that these data are for C_{12}

b. Forwards the information to Registry 1 after it receives $H(R_{+2}, RQ_{12})$ and $H(R_{+2}, ID_j)$ from Registry 2, and tells Registry 1 that these data are for C_{12}

2) Both Aggregator 1 and Aggregator 2 will do the same thing for the other hash values.

5. Each registry rehashes the values from aggregators.

1) Registry 1 re-hashes the values after it receives $H(R_{+2}, RQ_{12})$ and $H(R_{+2}, ID_j)$ from Aggregator 1 and sends them back as $H(R_{1+}, H(R_{+2}, RQ_{12}))$ and $H(R_{1+}, H(R_{+2}, ID_j))$.

2) Registry 2 re-hashes the values after it receives $H(R_{1+}, RQ_{12})$ and $H(R_{1+}, ID_i)$ from Aggregator 1 and sends them back as $H(R_{+2}, H(R_{1+}, RQ_{12}))$ and $H(R_{+2}, H(R_{1+}, ID_i))$.

3) Each registry will do the same thing for the other hash values.

6. Each aggregator verifies the data sources and performs matching.

1) Each aggregator

a. Check if $H(R_{1+}, H(R_{+2}, RQ_{12})) = H(R_{+2}, H(R_{1+}, RQ_{12}))$. If so, the data are from valid registries, and continue the remaining steps; otherwise the data are from invalid registries and the data cannot be used.

b. Check if $H(R_{1+}, H(R_{+2}, ID_j)) = H(R_{+2}, H(R_{1+}, ID_i))$. If so, then the same patient exists in Registry 1 and Registry 2.

c. Determine the number of matching identifier hashes in both registries. Let N_1 be the number of matched patients in Aggregator 1; N_2 be the number of matched patients in Aggregator 2.

2) Aggregator 1 determines the list of hashes to be re-conciliated.

- Let X_1 = {x | x is the hash value from Registry 1 who is not matched in Aggregator 1}
- Let Y_1 = {y | y is the hash value from Registry 2 who is not matched in Aggregator 1}

3) Aggregator 2 determines the list of hashes to be re-conciliated.

- Let X_2 = {x | x is the hash value from Registry 1 who is not matched in Aggregator 2}
- Let Y_2 = {y | y is the hash value from Registry 2 who is not matched in Aggregator 2}

7. Two aggregators reconcile the patients and send the partial counts to the data integration service if the counts are allowed to be disclosed.

1) Two aggregators
   a. Run a secure two-party addition protocol (Samet et al 2009) for each x in X_1 and y in Y_2. Assume Aggregator 1 initiates the protocol, and there are total M_1 matching patients in the end.
   b. Run a secure two-party addition protocol (Samet et al 2009) for each x in X_2 and y in Y_1. Assume Aggregator 2 initiates the protocol, and there are total M_2 matching patients in the end.
2) Aggregator 1
   a. Compute $A_{12}^1 = N_1 + M_1$.
   b. Send to Data Integration Service
3) Aggregator 2

a. Compute $A_{12}^2 = N_2 + M_2$.

b. Send to Data Integration Service

8. The data integration service calculates the final counts.

1) Calculate $C_{12} = A_{12}^1 + A_{12}^2$ after receiving A_{12}^1 and A_{12}^2.

2) Calculate other counts C_{11}, C_{21}, C_{22} using the same steps as C_{12}.

Note: one needs to be careful to address the "small cell" problem (where there is a risk of re-identification when aggregate values are small) in step 7 above, points 2 and 3, when the aggregators send the counts to the data integration service. In this case, we may need to restrict the aggregators from sending their counts directly to the data integration service. The two aggregators obtain the partial counts and then jointly compute the statistics that are required for the data integration service such as chi-square, odds ratio and relative risk. This can be addressed by extending the protocol as described in (El Emam et al, 2010, under review).

4.3.3 Security Analysis

The secure computation protocol adopts one-way accumulator and two-party addition protocol to achieve privacy and security during linking and computation. Table 4-3 is a summary of how the protocol meets the requirements identified in Section 4.3.1.

Table 4-3 Security analysis for multiparty secure computation linking

#	Requirement Description	Protocol Description
1	Able to link vertically partitioned data and calculate the aggregates across multiple registries	One way accumulator is used for matching identity; the aggregates are computed by calculating the number of the matched identity.
2	Protect the identity of patients	Identifier of patient is hashed before submitting, and each provider generates different hash value for the same patient.
3	Handle re-identification risk for population-based reporting	Only aggregates (counts) or statistics are disclosed.
4	Prevent any single third party or adversarial party to know with certainty the true data for a patient	Semi-trusted model. Two aggregators. No party can know the true value.
5	Be trustworthy and ensure that the data comes from valid registries	One way accumulator is used for membership testing.

Chapter 5. A Systematic Approach to Protect Privacy

Chapter 4 introduced three new privacy preserving data integration protocols for each of the three types of public health surveillance we identified in chapter 3. However, privacy-preserving data integration protocols cannot be successfully implemented in practice without a supporting organizational framework. In this chapter, we present a framework to assist in the implementation of privacy-preserving data integration protocols that provides a context for implementation in terms of identity protection, architecture, and methodology.

5.1. Framework Overview

Figure 5-1 shows the overview of the proposed framework. Methodology, architecture, identity protection and protocols work together to support the development of privacy preserving data integration for public health surveillance. When conducting a data integration project, a proper data integration protocol must be selected to meet the requirements of a specific health surveillance; identity protection mechanisms must be leveraged and addressed within organizations; and an appropriate architecture must be set up to enable integrating data cross organizations; finally a clear methodology must be followed to turn the protocol into practice. In particular, the methodology guides the data integration process that emphasizes privacy protection; the architecture defines the components that the trusted or semi-trusted third parties are used to enhance privacy; identity protection uses anonymization techniques or pseudonymous techniques; privacy preserving data integration protocols preserve the privacy when the data is transferred

and integrated among health care networks using specific algorithms within the context of methodology, architecture and identity protection.

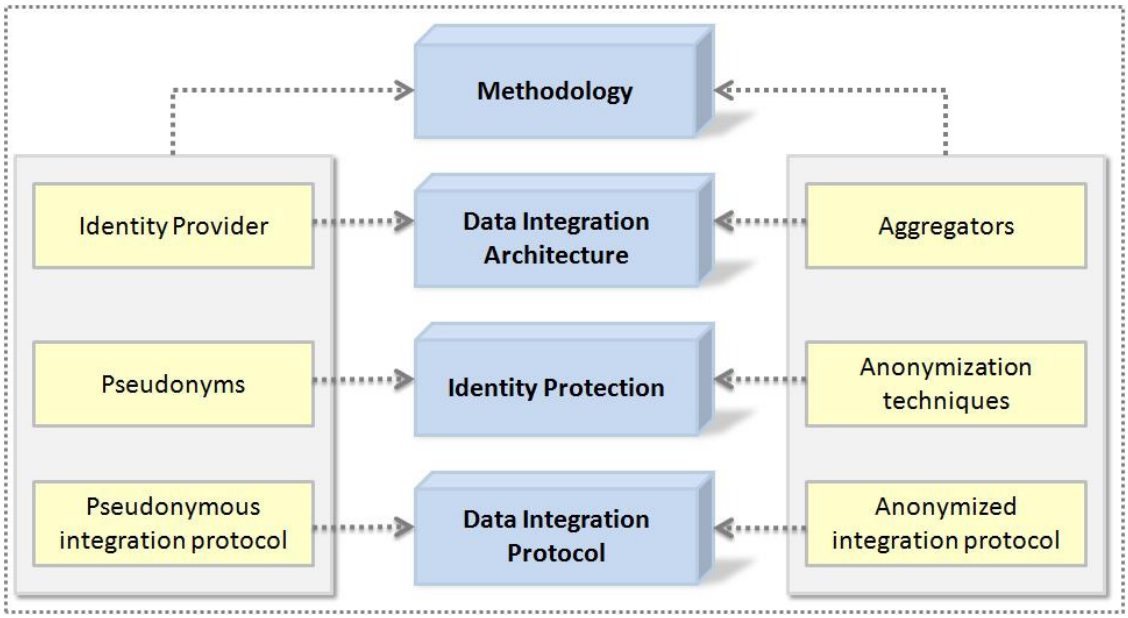

Figure 5-1 Privacy preserving data integration framework

5.2. Data Integration Protocol

The first step in implementing a protocol is to ensure you have selected one that is appropriate for the participating organizations and required public health surveillance. Chapter 6 has a detailed and systematic classification of data integration patterns that can be used to help in the selection. The levels of trust among all stakeholders and potential of re-identification risk are two important factors to decide which pattern and protocols to use. The trust relationship among the stakeholders, data providers, third parties, data integration service and clients, can be classified as the following trust levels: Trusted, Semi-trusted (as described in Section 2.4.1) and Untrusted; as shown in Table 5-1:

Table 5-1 Classification of trust levels

Level	Classification	Description
3	Trusted	The stakeholders are fully trusted to protect privacy and patient information can be shared with them.
2	Semi-trusted	The stakeholders are trusted to follow the rules. They will not act maliciously in an attempt to break the protocols used. However, patient information should not be shared with them.
1	Untrusted	Stakeholders cannot be trusted to follow the rules and protect patient information. They may act maliciously.

(El Emam 2010) defined the continuum of identifiability from low to high risk of re-identification: Aggregate data, Managed data, Exposed data, Masked data and Readily identifiable data. Based on these definitions, our framework provides a generic risk classification of re-identification of shared data and integrated for both patients and data providers. This classification is used to select proper data integration patterns and protocols for data integration. Table 5-2 shows the classification of potential risk of re-identification and their descriptions.

Table 5-2 Risk classification of re-identification

Level	Classification	Description
3	High	Dataset contains readily identifiable data of patients or data providers.
2	Middle	Dataset contains masked or exposed data of patients or data providers.
1	Low	Dataset contains managed or aggregated data of patients or data providers.

5.3. Identity Protection

The framework introduced by this thesis includes methodology, architecture, and protocols. With our three scenarios in section 3, we have identified three types of privacy-preserving data integration:

1. Aggregation
2. Pseudonymnous record linking
3. Anonymous record linking

As pointed out in Section 2.3.1 identity is central to privacy protection, and there are essentially two types of identity protection: anonymization and pseudononymization. We will treat aggregation as a special type of anonymization, since it keeps the patient identity hidden. Therefore, our framework has to balance and support both anonymized data integration and pseudonymous data integration. In addition, trusted or semi-trusted third parties are useful for privacy protection. Identity Provider provides federated pseudonymous identity management across organizations; Key holders are used for cryptosystem management during data integration; and Aggregators help calculate the aggregates and other statistics on either plain or encrypted data.

5.4. Architecture

Software architecture determines how best to partition a system, how components communicate with each other, how information is communicated, how system elements can evolve independently. The architecture described in Figure 5-2 shows the participating entities and security components.

5.4.1 Architecture Components

Three new data integration protocols introduced in Chapter 4 can be implemented based on a common architecture shown in Figure 5-2. The presence of the trusted/semi-trusted third parties and the data integration service is for the purpose of separating organization responsibility and enhancing privacy protection. The architecture is a Service Oriented Architecture in which all the Data Providers, the Identity Provider, Aggregator, the Key Holders and the Data Integration Service implement web services. While they coordinate and communicate via the standard web service interface, they also support specific API calls related to data integration protocols. The key components are described in detail below:

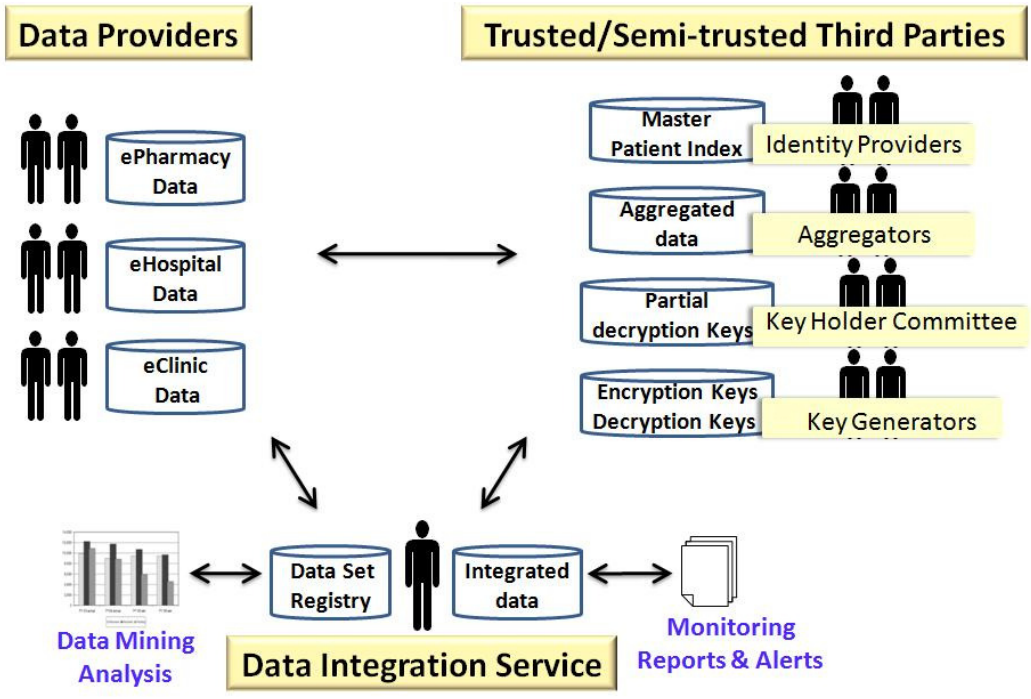

Figure 5-2 Privacy preserving data sharing and integration architecture

- **Data Provider** is a data source owned by health care organizations that publishes data for use by other organizations. It is responsible for the integrity and quality of the data it provides. Typically, wholesale access to its data is not provided, rather it must carefully define the datasets that can be requested. The datasets can be encoded or encrypted before delivery to enhance privacy and security. Patient identity can be protected by anonymization or aggregation, or a system of pseudonyms can be used in place of the actual patient identity. The Data Provider publishes its data by registering it with the **Data Registry** located in a **Data Integration Service**.
- **Data Integration Service** makes requests for datasets from **Data Providers** as defined and documented in the **Dataset Registry**. It links and integrates these datasets to make consolidated datasets in a privacy preserving way by working with other components (such as the trusted/semi-trusted third parties) according to carefully defined protocols and access control. It publishes the results as a service for external users for further analysis. Requests can only be made for registered datasets for which all necessary business and legal approvals are obtained. The Data Integration Service maintains a registry which controls who can have access to what datasets or reports and is responsible for monitoring and ensuring regulatory compliance.
- **Trusted/Semi-trusted Third Parties** are the third parties operated by some rules that enable trusted data integration by either providing federated identity management or offering comprehensive cryptosystem management. They are

considered to be semi-trusted parties and only trusted as long as they follow the rules.

- **Identity Provider** provides federated identity management. It provides single sign on for patients and health care workers. Encryption technology is used to create a Master Patient Index of pseudonyms for each person for each data provider and provide federated trust relationships with a high level of granularity, allowing the participating organizations to expose only a subset of their data; and enable an integrated, composite view of a single patient whose data is located in multiple and disparate data sources.
- **Aggregator** is a semi-trusted third party during data integration. It performs operations on the plain or encrypted data, and then sends the results to other components defined by the data integration protocol. In a simple aggregation, Aggregator performs the aggregation on plain data and obtains a clear aggregate. In a complex anonymized secure computation, one or more Aggregators perform the aggregation on ciphertext and obtain encrypted aggregates.
- **Key Holder Committee** is a group of the entities or servers that hold the keys for decrypting the encrypted data. Each member alone cannot recover the encrypted data.
- **Dataset Registry** registers the dataset to be collected and accessed across a B2B network. Access policy covers who can define, model and build the datasets, and who can access the datasets once they are built, by defining and running reports.
- **Data mining analysis, monitoring reports and alerts** are examples of results from public health surveillance or other knowledge discovery projects. For example, in

syndromic surveillance one might want to analyze trends or look for factors in the spread of disease, but also get continuously updated reports and charts, as well as alerts, if a new "outbreak" is detected in a new region.

5.4.2 Dataset Registry

Dataset Registry is a key component in our architecture. Much of this work was first introduced in (Hu et al, 2008).

Role of the Dataset Registry

The Dataset Registry records all dataset descriptions, relevant business agreements, consent forms and access control policies to allow the data collection and access associated with the dataset can take place. The role of the Dataset Registry is defined as:

- Establish a central point to register and manage datasets;
- Provide a complete technical specification of datasets;
- Serve as the system of record for dataset information throughout its lifecycle (Figure 5-3);
- Support the governance including access control to dataset information;
- Provide a standard, interoperable means for access, query and manipulation of datasets.

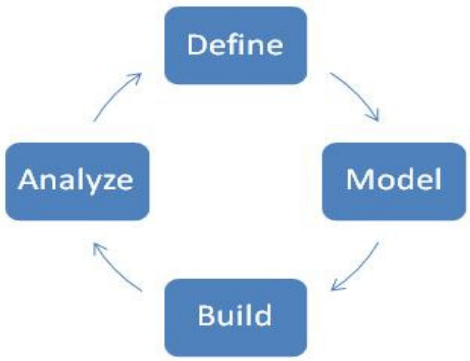

Figure 5-3 Dataset lifecycle

Dataset Registry handles the management of dataset descriptions and serves as the system of record for this information throughout the complete lifecycle of a dataset:

- **Define**. A dataset is defined in terms of the organizations that are the sources for the data used to build the dataset, and the business agreements between those organizations. Access control policies are then used to formally specify who can do what with the dataset that will be created.
- **Model**. A dataset is modeled in terms of joining together data extracts from data sources from each of the different organizations in the business agreement. Access control policies can be refined at this point, to control access down to the level of dataset attribute.
- **Build**. A dataset is built by the Data Integration Service. A dataset can be built once, or refreshed on an ongoing basis in order to keep the data current.
- **Analyze**. A dataset is analyzed by defining reports and queries that can be run against a dataset without affecting the data sources. Access control policies are created and updated to control access to reports.

The existing dataset definition will be updated if the analysis of the data set may lead to new requirements and new datasets will be created in order to extend public health surveillance. Throughout the entire lifecycle an audit trail of all changes and access to the dataset is maintained.

Access Control Model in Data Registry

Access control implemented in the Data Registry combines policy-based, role-based and team-based access control models to enhance consent and privacy protection. As shown in Figure 5-4, the permission is bound by business agreements and data set principles. This model has the following components:

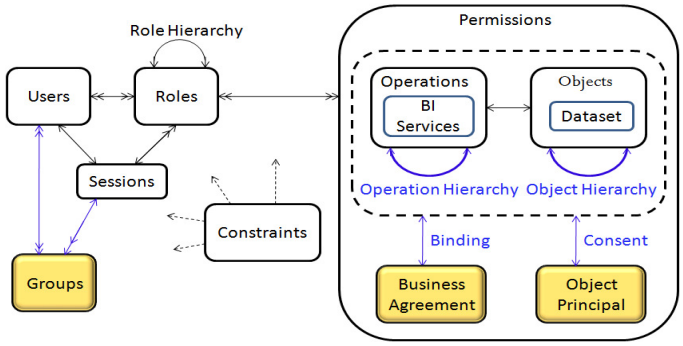

Figure 5-4 Consent-based Access Control Model

- **Entity**: User, Group, Role, Permission, Business Agreement, Object Principal, Constraint, Session
- **Assignment**: User-Role Assignment, User-Group Assignment, Group-Role Assignment, Role-Permission Assignment
- **Hierarchy**: Role Hierarchy, Operation Hierarchy, Object Hierarchy

Data Model of the Dataset Registry

The data model of the Dataset Registry is shown in Figure 5-5. *Dataset* and *Organization* are two main entity tables. *Dataset* is linked to the participating *Organizations* by *Business Agreements*. *Datasets* are made of *Extracts* that are defined in terms of *Columns.* These *Columns* comes from the *Views* that *Organizations* published for their *Data Sources*. *Reports* are built on top of *Datasets*.

Access control in the Data Registry implements the model by combining policy-based, role-based and team-based access control mechanisms. *Access Control Policy* table holds all policies which specify who (*Role*) can do what (*Permission*) to which (*Domain*) in what condition (*Constraint*). Whenever an action occurs, the policies are checked and a permit or deny decision is made. The description of this metadata is described as follows.

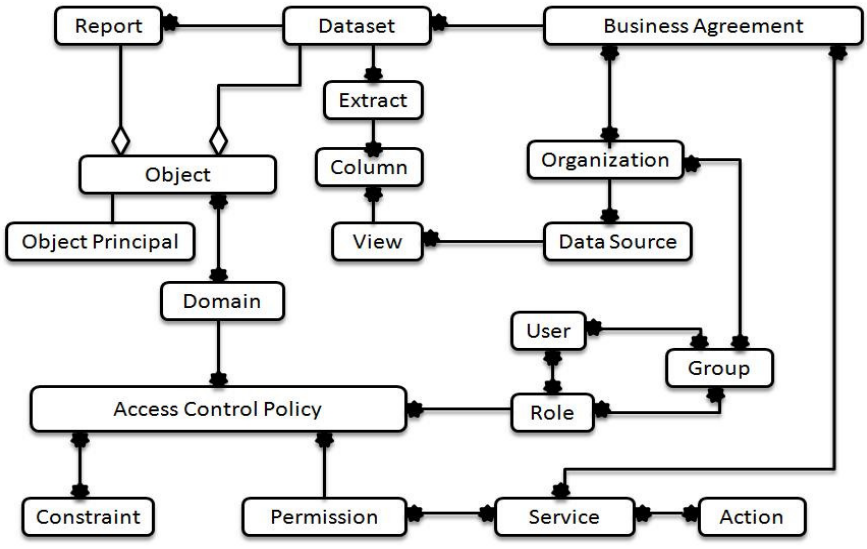

Figure 5-5 Data Model of the Dataset Registry (Hu et al, 2008)

- **Dataset** A set of data registered to be shared, accessed and collected.

- **Extract** A query request to extract data from a Data Source.
- **Column** A Column of a Data Source View for defining an attribute of a Dataset.
- **View** A logical table from a Data Source.
- **Report** A presentation of information queried from a Dataset.
- **Organization** An organizational entity.
- **Business Agreement** A contract between Organizations that regulates how data will be shared and accessed by who for what purposes.
- **Data Source** Data owned and maintained by an Organization.
- **Access Control Policy** Defines which role has permission with what constraint.
- **Permission** Stores all permissions about services.
- **Service** A group of one or more Actions that work together to complete a task, which bound with a business agreement.
- **Action** Stores all actions, including define, create, view, and execute access related to a Dataset or its Reports.
- **Role** A functional role that a User may be given with a single Organization.
- **User** Login id of anyone who can access a Dataset.
- **Group** A group of one or more users, roles and organizations that can be assigned a single role.
- **Object** A Dataset or a Report.
- **Object Principal** Entity that owns an object.
- **Domain** A group of one or more Object.

5.5. Methodology

In our proposed framework, we have extended and refined the CRISP-DM data mining methodology, to create our proposed privacy preserving methodology for public health surveillance (shown in Figure 5-6) to address the new requirements implied by data sharing in a B2B health care network. This methodology was first introduced in (Hu & Peyton, 2010). It is revisited in the context of a data integration architecture that allows for the continuous integration of data from multiple organizations. It has advantages in terms of essential elements such as platform, availability, data used, authorization, privacy protection, data sharing, data integration, modeling techniques, monitoring and security.

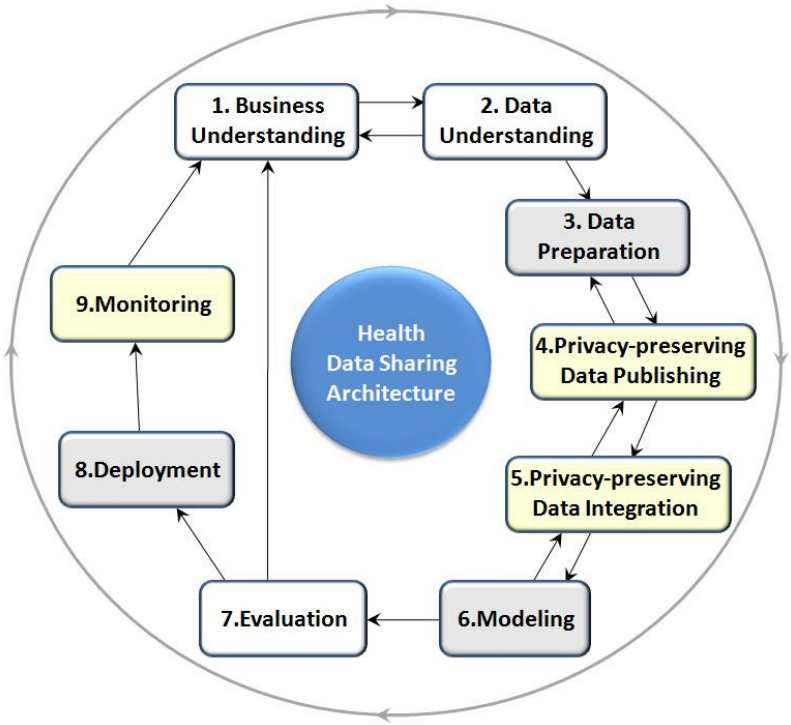

Figure 5-6 Privacy preserving methodology for public health surveillance
(Yellow = new steps, Grey = modified steps)

The proposed methodology consists of nine phases. As shown in Figure 5-4, we add and modify the phases based on CRISP-DM to address the special privacy requirements in a B2B health care environment. “4. Privacy-preserving Data Sharing”, “5. Privacy-preserving Data Integration”, and “9. Monitoring” are new phases to address the issues of trust, regulation compliance and data sharing across organizational boundaries. We also modify three other phases, “3. Data Preparation”, “6. Modeling” and “8. Deployment” to address new business models for public health surveillance. In the following section, we will discuss in detail how the privacy preserving requirement must be addressed at each phase, as well as where the appropriate techniques and technology are used.

1. Business Understanding

The first phase focuses on understanding the business requirements and the definition of a suitable and continuous public health surveillance process. There are two main requirements that must be addressed from a business point of view.

- Establishing the business relationships and agreements that allow data to be collected from different organizations;
- Ensuring that appropriate patient consents and/or organizational agreements, and privacy safeguards are in place.

For example, in the syndromic surveillance system described in Section 3.1, data needs to be collected from clinics, hospitals and other health practices across Canada in order to monitor key influenza indicators to determine the severity and spread of influenza, and detect trends in the spread of influenza. First the business relationships and agreements

must be established to allow aggregate influenza related data to be collected from different data providers and a proper organizational authorization should have been given to use the data in influenza surveillance. Because of privacy concerns, neither patients nor data providers want to be identified when the surveillance results are public. Therefore, privacy of patient and data provider should be protected during data collection and data integration.

2. Data understanding

The data understanding phase involves understanding

- What data must be collected from which data providers and what attributes can be used to link and integrate the data.
- How identity and privacy are to be safeguarded. There may be patient-defined as well as organization-defined restrictions on individual attributes collected from the data sources.
- How a common view is maintained and presented in a consistent fashion if the data is published.
- How the resulting attributes in the consolidated data set do not result in situations where patients become potentially identifiable or sensitive data can be inferred.

For example, in the syndromic surveillance system described in Section 3.1, it is sufficient for detecting meaningful trends to only collect the counts of patients presenting with a particular syndrome, such as gastro-intestinal symptoms (GI) and influenza like illness (ILI). Individual patient level data is not needed, and the total counts of ILI and GI cases are monitored by patient age groups and geographic groups. Because the

surveillance results are shown in a map, care must be taken to ensure that data providers are not potentially identifiable or be inferred.

3. Data Preparation

In addition to the tasks of selecting data, cleaning data, constructing data, integrating data, and formatting data used for a single organization defined in CRISP-DM, our framework needs each Data Provider to prepare data so that it can be published to make it accessible by other organizations. There are four main requirements specific to health care in this phase:

- Data standardization.
- Normalization
- Data cleaning.
- De-personalization.

For example, in the syndromic surveillance system described in Section 3.1, each data provider who participates in the surveillance will determine its patient counts for each stratum every 24 hours. Patient information is not included for submission to ensure privacy protection. And the counts are encrypted to address privacy of data provider. After the counts are encrypted by a given key, it is ready to be published and be used in a public health surveillance process.

4. Privacy-preserving Data Publishing

This step defines how and when the data can be made available. The four main requirements specific to a health care network that must be addressed in this phase are

- Patient consents and/or organizational agreements for sharing
- Standardization of publishing
- Security of the data
- Protection of privacy.

There are several kinds of data for publishing, which depends on the data integration requirements. It could be the aggregate data, encoded, encrypted or non-encrypted, when no patient level data is required. It could be the patient level data that has been de-identification, or masked, or encrypted. For example, in the syndromic surveillance system described in Section 3.1, each data provider registers with the Data Integration Service in order to participate. An application is used to allow the data provider to submit data every day and ensure that the dataset conforms to the common data view, i.e. a consistent handling of encrypted counts for GI and ILI. The communication between the data provider service and the aggregator service is a standard web service interface.

5. Privacy-preserving Data Integration

The Privacy-preserving Data Integration phase is introduced to integrate data sets from different Data Providers, which may be arriving continuously and asynchronously. This is the critical phase where the technical approach for enabling privacy-preserving data integration to create integrated data sets that can be used as a basis for public health surveillance across the B2B network. The Data Integration Service, through its interaction with the Identity Provider requests the data sets it desires from each Data Provider. Master Patient Index in the Identity Provider plays a key role when linking identities across organizations. It is important that the Master Patient Index is kept

separate from the actual data otherwise one would be able to link all health data in the network to the real identities.

Once the technical approach to enabling privacy-enabled data integration is selected, the basic steps to realize the approach are as follows (Figure 5-7).

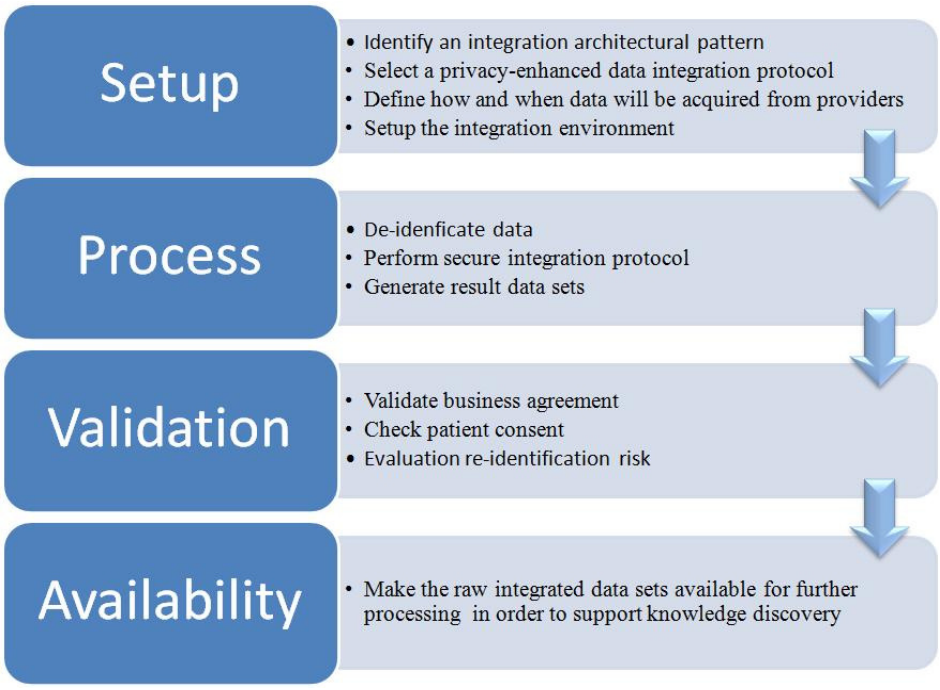

Figure 5-7 Data integration process

Step 1: Setup of Data Integration Environment. This is a step to identify architecture, protocol, initialize and set up the integration environment. First, an architectural pattern can be identified to best meet the requirements. Next a privacy preserving data integration protocol is selected to be compatible with the identified architectural pattern. After defining when and how data will be acquired from data providers, a data integration environment should be setup. For example, if encryption is required, trusted or semi-trusted third parties are needed; and encryption and decryption keys are needed to be

generated and delivered. For example, in the syndromic surveillance system described in Section 3.1, k-key holder provider anonymized data aggregation protocol is chosen.

Step 2: Processing of Data Integration. This step is actually performing data integration operations and process, and generation of datasets. Based on the selected data integration protocol, collected data is re-identitfied, and integration operations are executed. Finally integrated datasets are produced for further analysis.

Step 3: Validation of Datasets. This step is to determine if the privacy is protected in the process of data integration and in the datasets that is generated by Data Integration. The following issues must be addressed:

- Appropriate consent is obtained from patients before collecting their data.
- Business agreements are in place both for obtaining the data from participant organizations and sharing the results.
- The individuals or services that will access the integrated data set have the required access rights. In addition, a threat risk assessment should be made to ensure that security is in place to prevent hacking or other forms of unauthorized access.
- The integrated dataset will not create identifiable data and the risk of re-identifying in combination with other sources of data must be minimized.

Step 4: Availability of Datasets. This is a step to make the raw datasets available to analysts and others according to consent, contracts and purposes. In particular, this includes further processing in order to systematically support public health surveillance and deploy the results.

6. Modeling

Modeling is used to structure, organize, filter, process the datasets into a form that facilitates and controls public health surveillance. In the modeling phase, depending on the specific public health surveillance task, various techniques are selected and applied, and their parameters are calibrated to optimal values. This phase considers various models and chooses the best one based on their predictive performance. There are two main issues specific to health care.

- Reduce potentially re-identification risk for the consolidated data set.
- Choose a proper modeling technique such as data mining algorithms, statistical techniques, OLAP data models, text mining and event or data stream mining to handle processing of large quantities of data arriving on a continuous basis.

In the syndromic surveillance system described in Section 3.1, statistical techniques are used. Totals and rates are calculated on the distribution of GI and IPI to show influenza trends with other reports generated upon the raw datasets.

7. Evaluation

Evaluation is the process of determining:

- If public surveillance has meaning in the targeted business scenario;
- If there still exists some important business issues that need to be further considered;

Evaluation criteria that can be used to measure success include quality of model, cost, security and privacy. In most health scenarios, the following should be verified:

- The results of public health surveillance process are accurate and timely;
- The performance of data integration is reasonable and is scalable;
- Consent and business agreement are in place;
- Access rights are applied.
- Re-identification risk is reduced to minimum.

8. Deployment

The results of the health surveillance process are published by the Data Integration Service and registered in its dataset registry. There are two key points in our methodology:

- The Data Integration Service is itself a Data Provider and the results of one public health surveillance process can be a data source that is used by another health surveillance process.
- All the usual constraints and process that apply to a Data Provider during the privacy-preserving data publishing phase, apply to the Data Integration Service when it deploys its surveillance process and publishes the results.

9. Monitoring

A new phase, "Monitoring", is added in our methodology to ensure regulatory compliance and ensure the quality and consistency of the data, as well as the availability and reliability of the Data Provider. Choosing an effective monitoring mechanism is vital for completing a successful public health surveillance project. (Baron et al, 2003) focus on the monitoring of patterns and the detection of interesting changes. (Peyton et al,

2007) proposes an audit trial service for verifying compliance with privacy regulations in a Circle of Trust for eHealth.

Chapter 6. Privacy Preserving Data Integration Patterns

In chapter 5, we presented our framework in which privacy preserving data integration protocols can be enacted, and in chapter 4, we described in detail our 3 new proposed protocols for public health surveillance. There are actually a wide variety of protocols in place around the world for public health surveillance. The choice of protocol is dependant both on the sophistication of the organizations involved in terms of methodology for managing surveillance and the information technology infrastructure and architecture available. It is also dependant, to a large extent on the tolerance for risk in protecting privacy and the willingness to trust other organizations.

In this chapter we present a set of privacy preserving data integration patterns for privacy preserving data integration. The set of patterns is representative of the types of protocols we have encountered in the literature as well as in our interactions with public health organizations, but it is not intended to be a complete set of patterns. Rather it is a tool that can help public health organizations assess and understand the spectrum of protocols that is possible, while guiding them to those patterns (like the three we propose) that are most optimized for providing integration while preserving privacy.

The patterns bundle protocols and techniques to address different types and levels of PHI protection (anonoymized vs federated pseudonyms) for different types of data integration (aggregated vs identity linking) according to the level of acceptable risk and trust involved.

6.1. Pattern Classification

A privacy preserving data integration pattern is a solution to a recurring privacy problem when integrating horizontally or vertically partitioned data cross organizations. Figure 6.1 is a pattern classification tree which gives an overview of privacy preserving data integration patterns. The patterns in grey are existing patterns that match protocols currently used in practice. Typically these require a trusted third party. The patterns in yellow and green correspond to the new protocols we proposed in Section 5. The green patterns are more secure with a semi-trusted model and with minimal risk of re-identification, while the yellow pattern still requires a trusted third party. The proposed protocols are being expressed in a pattern form but they are preliminary and have only been used in our case studies so far.

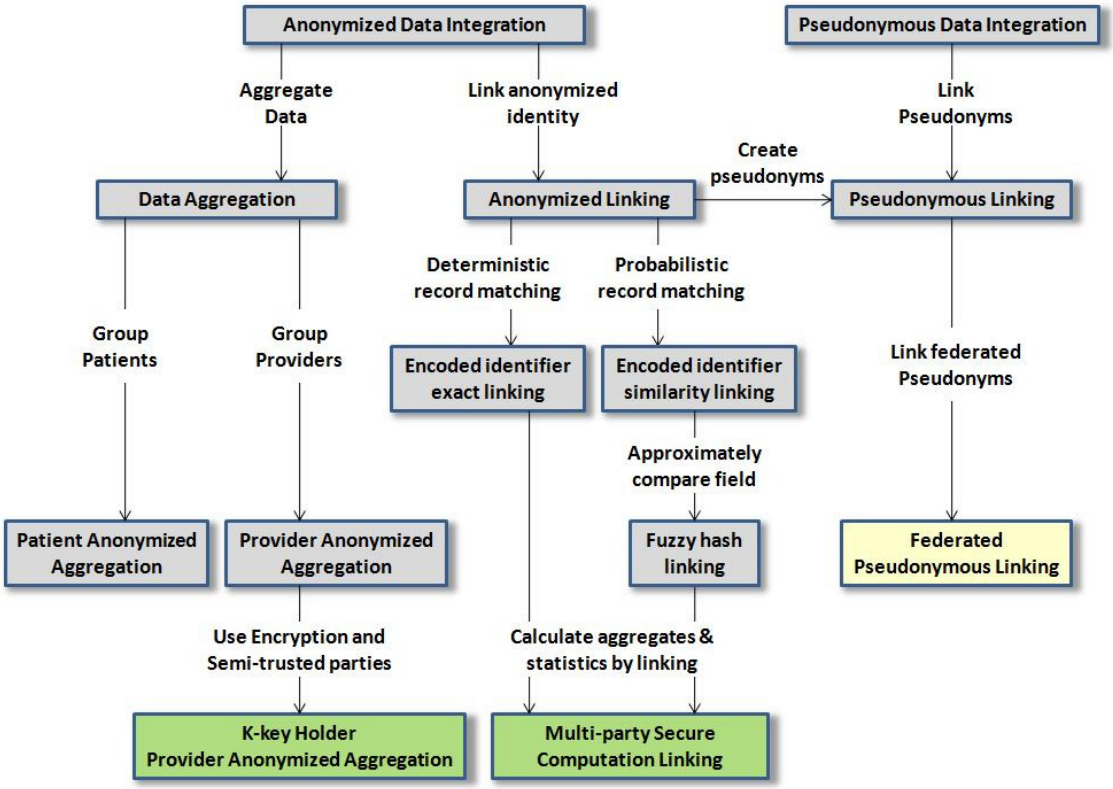

Figure 6-1Privacy preserving data integration patterns

As shown in Figure 6-1, the initial high-level classification of patterns is into two types of privacy-preserving data integration for public health surveillance: “Anonymized Data Integration” and “Pseudonymous Data Integration”. We consider “Data Aggregation” to be a special type of “Anonymized Data Integration”. Our three examples in chapter 3, correspond to this initial high level classification (3.1.3 Data linking and aggregation, 3.1.2 Record linking, and 3.1.2 Data aggregation).

The “Anonymized Data Integration” pattern describes how to integrate and protect sensitive data from multiple data sources by anonymizing data. Generally, data should be anonymized before it is sent for integration. There are many different types of anonymization techniques that can be used. “Data Aggregation” and “Anonymized linking” are commonly used in public health surveillance cases. For “Data Aggregation”, anonymization techniques can be used to either protect the identity of the patient or the identity of the provider of the data, or both. For “Anonymized Linking” there are two linking strategies. The “Encoded identifier exact linking” pattern uses a deterministic matching method where the fields of an identifier are encoded and evaluated by equivalent testing. The “Encoded identifier similarity matching” pattern uses a probabilistic matching method where the similarity techniques are leveraged to calculate the similarity of two identifiers.

“Pseudonymous Data Integration” describes how to integrate, link and protect sensitive data from multiple data sources through the use of pseudonyms created by

trusted third party. Generally a trusted third party is used to do pseudonym mapping, and any privacy preserving data linking techniques can be used for creating pseudonyms.

6.1.1 Principles for Characterizing the Patterns

The privacy preserving data integration patterns are characterized based on the following principles of privacy protection.

1) **Semi-trusted vs trusted third party**. Using semi-trusted third party is securer than using trusted third party since it is not necessary to fully trust the third party (leaving one vulnerable to single malicious third party attacks).
2) **Anonymized vs pseudonymous techniques**. Using anonymized techniques protects identity of patients and providers without tracing back the protected identities. Pseudonyms protect identity of patients and providers as well as allow tracing back to the protected identities (with the cooperation of the Identity Provider).
3) **Encrypted vs not encrypted identity information**. Encryption helps protect confidentiality and integrity of data.
4) **Separated data**. Separating sensitive data such as identity information from other data such as health information enhances privacy protection.
5) **Partial data**. Holding only partial data in different parties prevents any single party from compromising privacy.
6) **Federated pseudonyms**. With federated pseudonyms, one can link events across sessions to a pseudonym identity without knowing the actual identity; and each organization has different pseudonym for an actual identity.

7) **Provider protection**. It should not only protect privacy of patients but also privacy of data providers.

In the following sections, the more specific patterns in the pattern classification tree will be described in detail. In particular, we will diagram the patterns in terms of the components they use from the architecture diagram in figure 5-2.

6.2. Data Aggregation

In this section, we describe a set of data aggregation patterns that apply different anonymization techniques to remove patient and data provider identity. Ideally, no intermediate information should be disclosed apart from the final aggregated results.

6.2.1 Pattern: Patient Anonymized Data Aggregation

The Patient anonymized data aggregation pattern describes how to integrate and protect sensitive data from multiple data sources by aggregating patient data.

Context:

Public health surveillance wants to detect unusual trends by collecting and analyzing the patient data that is located in different health care organizations. Aggregated data is useful for such analysis.

Problem:

How to collect and aggregate patient data from multiple distributed organizations while ensuring patient identity is protected?

Forces:

1) Patient data is sensitive.

2) Patient identity information cannot be disclosed.

3) Privacy legislation requires protecting personal data when collecting and using it.

4) There is not a single identifier.

5) Patient level data is not required.

6) Aggregated counts of patients grouped by some sub-populations are useful.

Solution:

The patient anonymized data aggregation pattern can be used in some public health surveillance scenarios when no detailed patient data is required. As shown in Figure 6-2, each of the Data Providers creates the aggregates based on patient groups, for example age groups 0-10, 11-20, … and sends them to the Data Integration Service, which combines and consolidates the aggregates into the integrated data repository. Note that the Data Providers and Data Integration Service are the same entities mentioned in the architecture in Figure 5-2.

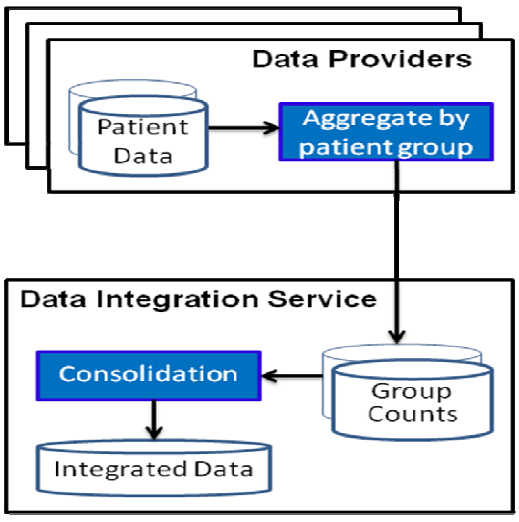

Figure 6-2 Pattern: patient anonymized data aggregation

Figure 6-3 shows the details of a specific example protocol based on the patient anonymized data aggregation pattern.

1) Each Data Provider calculates aggregates by patient group. In this way, patient's identity is removed and patient data is anonymized.
2) Each Data Provider sends group aggregates to the Data Integration Service.
3) The Data Integration Service manipulates the aggregates from all Data Providers. It is unable to relate any of the data to a specific patient.

However, aggregation or other statistical techniques for anonymization are not immune to re-identification risk. Moreover, this protocol does not address protecting the identity of the data provider.

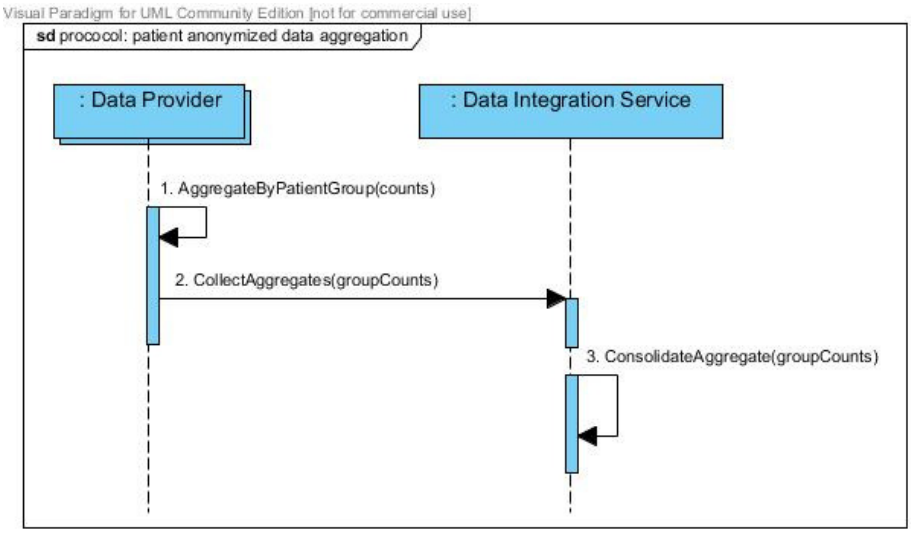

Figure 6-3 Protocol: patient anonymized data aggregation

6.2.2 Pattern: Provider Anonymized Aggregation

Provider anonymized data aggregation pattern describes how to aggregate data from multiple data providers while protecting the identity of data providers by using one or more regulated third parties.

Context:

Public health surveillance wants to collect data from different health care providers for analysis or disease surveillance, where individual patient level data is generally not needed to detect epidemics, but some demographic and other information are helpful to detect and track epidemics.

Problem:

How to collect and aggregate patient data from multiple distributed organizations while addressing privacy of patient and protecting identity of data provider?

Forces:

1) Patient data is sensitive.
2) Privacy legislations control personal data collection.
3) Data providers are concerned about their privacy and risk of reputation management, and may be unwilling to share data.
4) Patient level data is not required.
5) Aggregated counts of patients grouped by some sub-populations are useful.
6) Grouping of data providers is required and useful for reporting, such as location based reporting.

7) Risk of re-identification is high. For example, if location based reporting is involved.

Solution:

The provider anonymized data aggregation pattern describes how to integrate data from multiple data providers while protecting the privacy of data providers by aggregation using a trusted third party - Aggregator. Aggregation is a technique for anonymizing patient data; it can be also used to anonymize data from a specific data provider. As shown in Figure 6-4, an aggregator is used to anonymize data providers for aggregation style data integration. Note that Data Providers, Aggregators and Data Integration Service are the same entities mentioned in the architecture in Figure 5-2. Each data provider calculates patient counts by patient group, and then sends these counts to the aggregator. The aggregator computes the sums by data provider group, and then sends to Data Integration Service. Data Integration Service generates integrated data and can be used by Public Health Unit for further analysis. Data Integration Service and public health can only know aggregated data and no clue to know true value of each data provider.

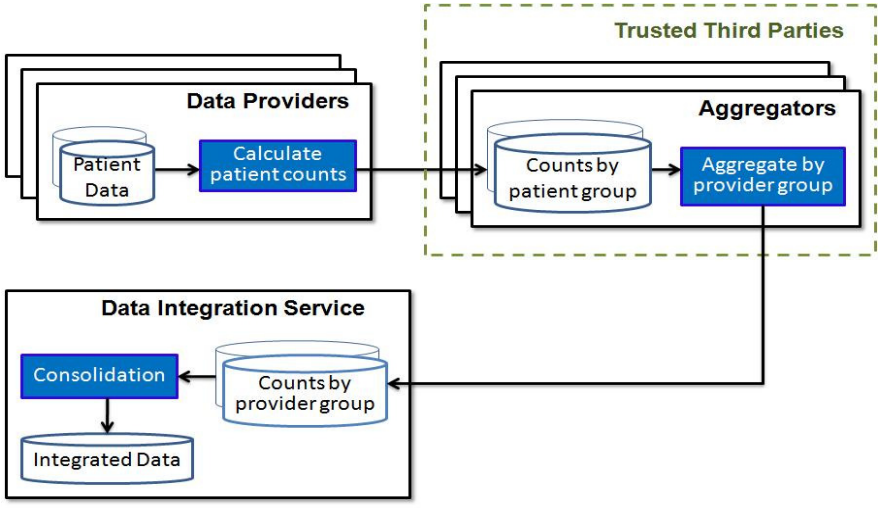

Figure 6-4 Pattern: provider anonymized data aggregation

One example is, in disease surveillance, each data provider determines its case and patient counts for each stratum and submits to public health authorities. The aggregate totals are collected at a few levels: local, province and national or even international health authorities. In this case, the local and province health authorities can serve as the trusted third party. The protocol for anonymized data aggregation by trusted third party is shown in Figure 6-5.

1) Each Data Provider calculates patient counts so that patient is protected.

2) Each Data Provider sends its aggregates to its aggregator.

3) Each aggregator calculates counts by provider group so that providers are anonymized.

4) Each aggregator sends its aggregates to Data Integration Service.

5) Data Integration Service consolidates counts from all Data Providers.

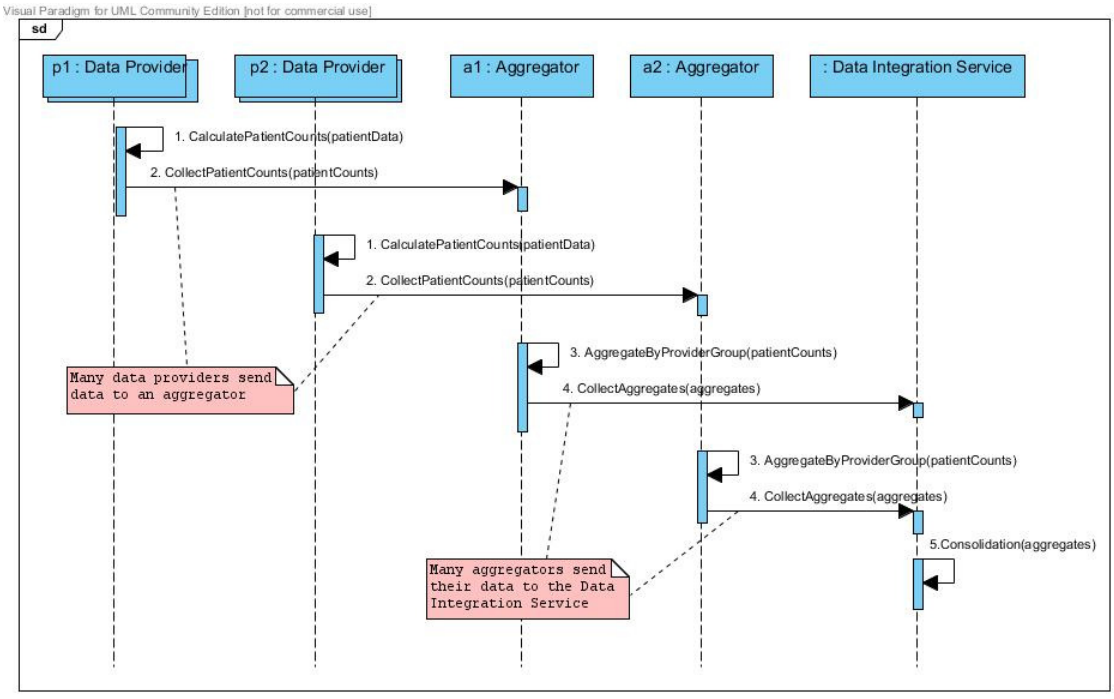

Figure 6-5 Protocol: provider anonymized data aggregation

This protocol assumes that Aggregators are fully trusted, and they won't reveal the real data from each data provider. However this is not ideal in terms of minimizing risk. Ideally, we should not depend on Aggregators to be trusted. Once the data providers disclose the data related to themselves or patients, they cannot control the usage of the data and their privacy is compromised. Therefore the data providers are not willing to submit data if they feel insecure.

6.2.3 Pattern: K-Key Holder Provider Anonymized Aggregation

The K-key holder provider anonymized aggregation pattern integrates encrypted aggregated data from multiple data providers using one or more regulated third parties to achieve maximums protection of the identity of providers. This pattern captures the essential elements of the protocol described in section 4.1.

Context:

Public health surveillance wants to collect data from different health care providers for analysis or disease surveillance, where individual patient level data is generally not needed to detect epidemics, but some demographic and other information is helpful to detect and track epidemics.

Problem:

How to collect and aggregate patient data from multiple distributed organizations while addressing privacy of patient and protecting identity of data provider to the maximum?

Forces:

1) Patient data is sensitive.

2) Privacy legislations control personal data collection.
3) Data providers are concerned about their privacy and risk of reputation management, and require maximum protection of their identities.
4) Patient level data is not required.
5) Aggregated counts of patients grouped by some sub-populations are useful.
6) Grouping of data providers is required and useful for reporting, such as location based reporting.
7) Risk of re-identification is high. For example, the location based reporting is involved.
8) Not using a trusted third party. Semi-trusted third parties are required.

Solution

The k-key holder pattern aggregates data from multiple data providers while protecting privacy of data providers by the use of a secure multiparty computation protocol. This addresses the problem that third party aggregators cannot always be trusted in the real situation, as noted for the prior protocol.

The concept of secure multiparty computation can be leveraged to learn the global results without revealing the real data at individual organizations. As shown in Figure 6-6, an aggregator and multiple key holders are used to help anonymize data providers and enable data integration. Note that Data Providers, Aggregators, Key holders and Data Integration Service are the same entities mentioned in the architecture as Figure 5-2. Each data provider calculates patient counts by patient group, and encrypts them, and then sends these encrypted counts to the aggregator. The aggregator processes the encrypted data, and computes the sums by data provider group, and then sends to each key holder.

The aggregator does not know the original values from each provider and does not know the sums since they all are encrypted. Each key holder decrypts the sums it receives and sends that value to Data Integration Service. Each key holder has no clue about the true values of the sums since they are only partially decrypted. Through combination process of partial decryption, Data Integration Service generates integrated clear data and can be used by its consumers for further analysis.

This pattern protects against the possibility of some parties colluding with each other. Especially it prevents any single adversarial party to know with certainty the true values for any data provider. It is also resistant to collusion between any aggregator and any one key holder. An example protocol for anonymized data aggregation through secure multiparty computation for protecting patient and data provider's privacy is shown in Figure 6-7. The protocol we presented in section 4.1 is an example of a protocol that fits this pattern. It is based on the additive homomorphic property and threshold version of a homomorphic cryptosystem such as Paillier (Paillier, 1999).

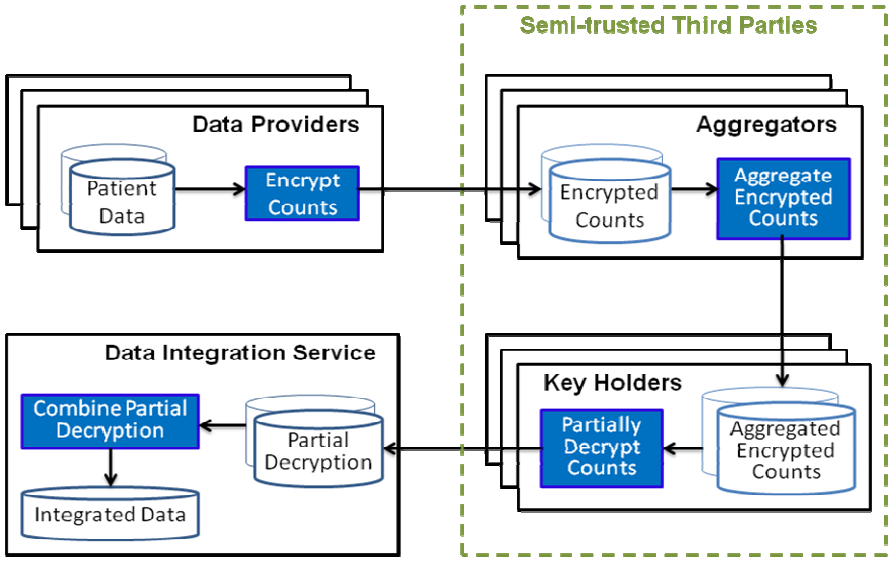

Figure 6-6 Pattern: K-Key Holder provider anonymized data aggregation

1) Setup: a key generator generates keys and delivers them to each Data Provider and three key holders.

2) Each Data Provider encrypts counts so that patient identity is protected.

3) Each Data Provider sends its encrypted counts to its aggregator(s).

4) Each aggregator sums encrypted counts.

5) Each aggregator sends its summation to each of three key holders.

6) Each key holder performs partial decryption through decryption algorithm.

7) Each key holder sends partial decryption to Data Integration Service

8) Data Integration Service consolidates counts from all Data Providers.

When implementing this protocol, it is crucial for each party to follow pre-defined guidance and protocol for processing data.

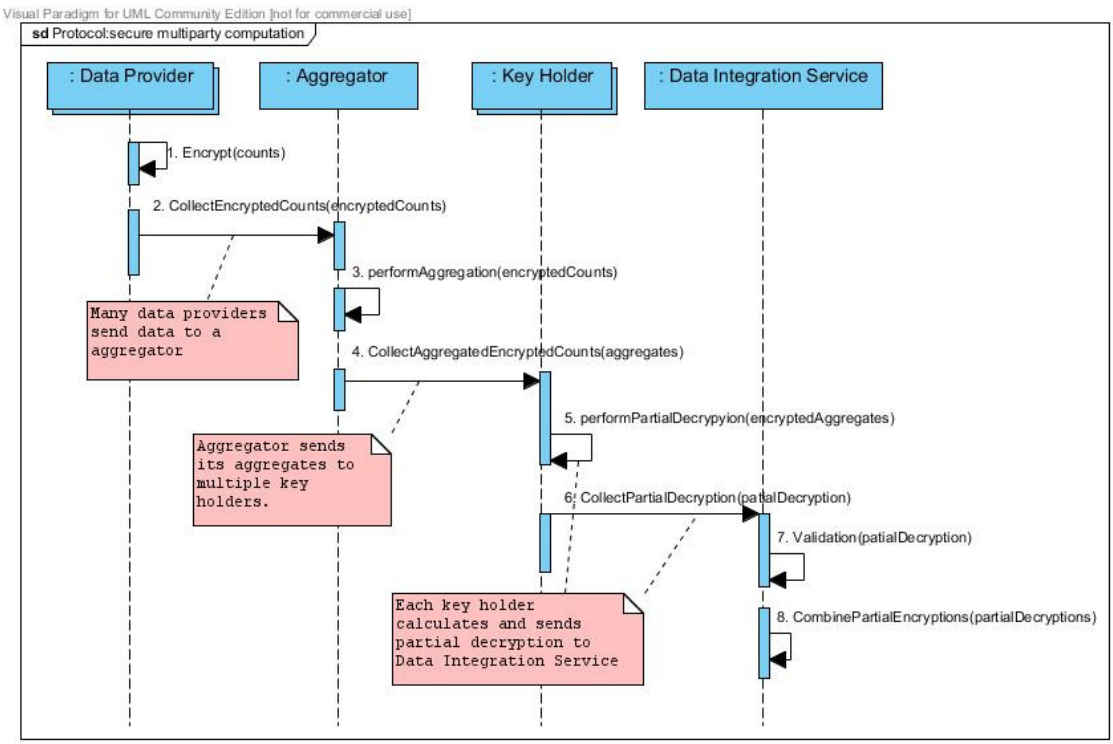

Figure 6-7 Protocol: K-key holder provider anonymized data aggregation

6.3. Pseudonymous Data Integration

Data integration is pseudonymous if one can link events to an identity created for a specific organization, without knowing the actual identity. Pseudonymous data integration consolidates data and links identity via pseudonyms without compromising identity. In this section, we analyze in detail the creation of pseudonyms, and the patterns, and protocols for pseudonymous data integration.

6.3.1 Pattern: Pseudonymous Data Federation

Pseudonymous data federation describes how to integrate, link and protect sensitive data from multiple data sources through the use of pseudonyms created by a trusted third party.

Context:

Public health surveillance wants to analyze patient data to find data relationships and detect unusual trends by collecting and linking patient data, where data for a patient is located in multiple, disparate data providers in a B2B health care network.

Problem:

How to create an integrated, composite view of a single patient while ensuring patient identity is protected?

Forces:

1) Patient data is sensitive.

2) Privacy legislations require protecting personal data when collecting and using it.

3) Patient identity information cannot be disclosed.

4) Linking identity is required.

5) Reduce re-identification risk.

6) There is a unique identifier.

Solution:

The pseudonymous data federation pattern allows linking patient data from multiple data providers using pseudonyms to protect patient identity. In particular, the pattern can be used when the organizations have the same patient identity attributes such as health card number. Patient identifier and health data should be encrypted before it is sent to the regulated third party for pseudonymization. As shown in Figure 6-8, the Data Provider encrypts the patient's identifier and patient health data separately. Then it sends them to the trusted third party. The trusted party performs pseudonymization, and sends the pseudonyms and encrypted health data to the Data Integration Service. Finally Data Integration Service performs integration process upon pseudonymous data.

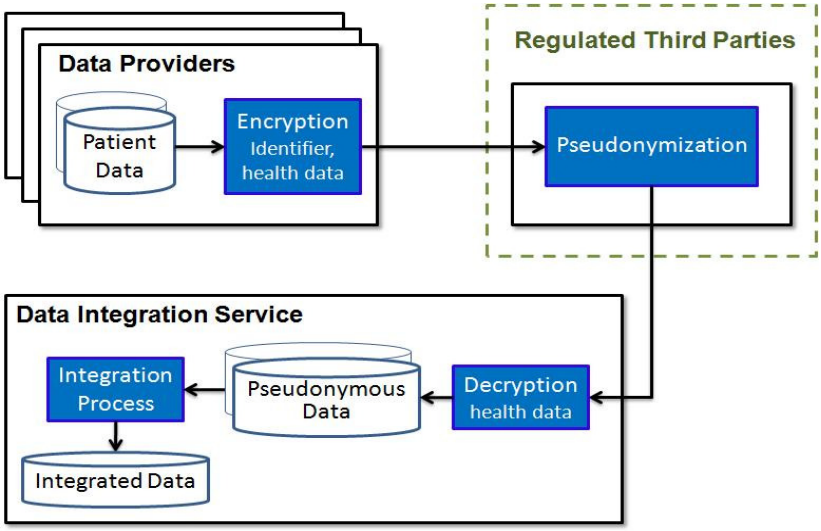

Figure 6-8 Protocol: pseudonymous data federation

The main issue for this protocol is that it requires a common identifier, and each patient has a single pseudonym across all data providers that raises privacy risks. An example pseudonymous data federation protocol is described as follows and shown in Figure 6-9.

1) User requests for a consolidated data set.
2) Data Integration Service transfers query into sub query.
3) Data Integration Service sends sub query to each Data Provider.
4) Each Data Provider performs data creates partial results.
5) Each Data Provider encrypted identifier and partial results separately.
6) Each Data Provider sends encrypted data to the regulated third party for pseudonymisation.
7) The regulated third party sends data to Data Integration Service with pseudonyms and partial results after pseudonymisation. Each patient has the same pseudonym.
8) Data Integration Service decrypts partial results.
9) Data Integration Service performs transformation process.
10) Data Integration Service consolidates data from all Data Providers.

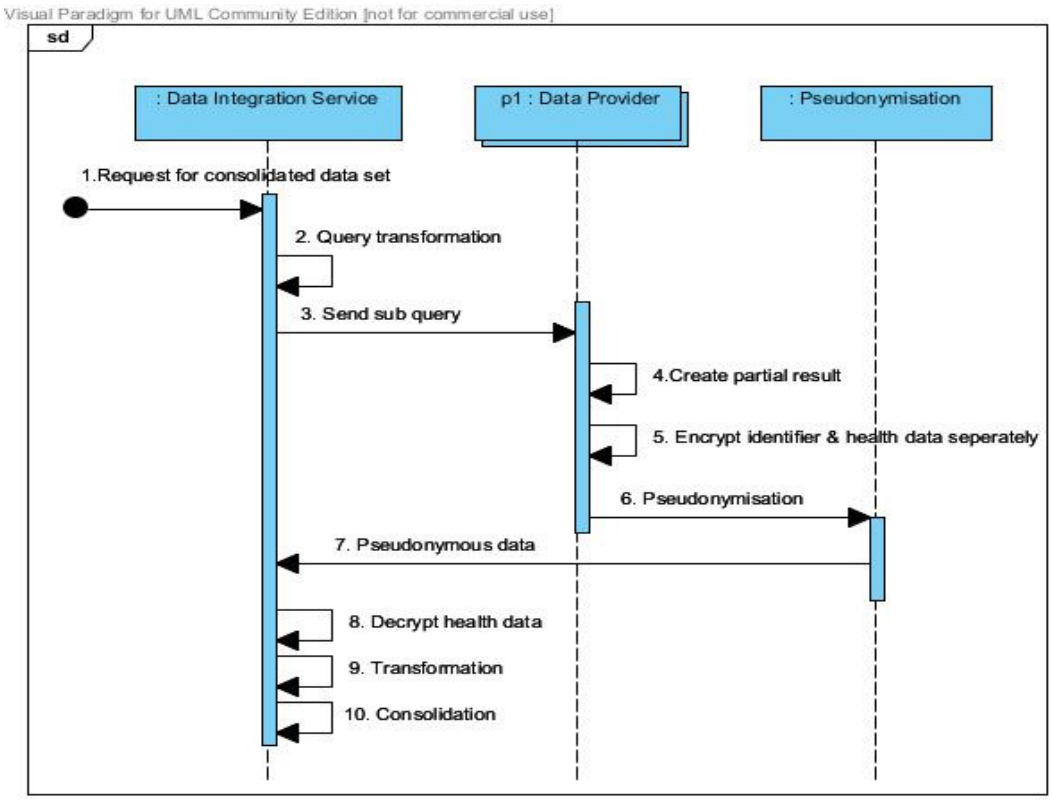

Figure 6-9 Protocol detail: pseudonymous data federation

6.3.2 Pattern: Master Patient Index Linking

Master patient index linking integrates, links and protects sensitive data from multiple data sources through the use of pseudonyms created by a trusted third party. This pattern captures the essential elements of the protocol described in section 4.2.

Context:

Public health surveillance wants to analyze patient data to find data relationships and detect unusual trends by collecting and linking patient data, where data for a patient is located in multiple, disparate data providers in a B2B health care network.

Problem:

How to create an integrated, composite view of a single patient while ensuring patient identity is protected?

Forces:

1) Patient data is sensitive.
2) Privacy legislation requires protecting personal data when collecting and using it.
3) Patient identity information cannot be disclosed.
4) Linking identity is required
5) Reduce re-identification risk
6) There is not a single identifier.
7) Identity data standardization issues.
1) Providers are willing to share data based on patient identity.
2) The linking mechanism is a reusable to improve efficiency.

Solution:

Using an Identity Provider and Master Patient Index, the same patient has different pseudonyms in different organizations. Federated identity management reconciles the pseudonyms and aggregates data to develop a consolidated view of the patient or link federated pseudonyms to obtain a set of statistical data that can be used for performance management, knowledge discovery and etc. Master Patient Index linking does not need a common identifier and its use of federated pseudonyms prevents collusion among organizations. The Master Patient Index must have been setup before a data integration process can be conducted. As shown as Figure 6-10, each data provider has their own

pseudonym for each patient. During the data integration process, the data is separated into two parts. The health data is sent to Data Integration Service. The pseudonym is sent to the Identity Provider for mapping. Identity Provider sends Data Integration Service a pseudonym for Data Integration Service who performs data linking and generates consolidated data sets.

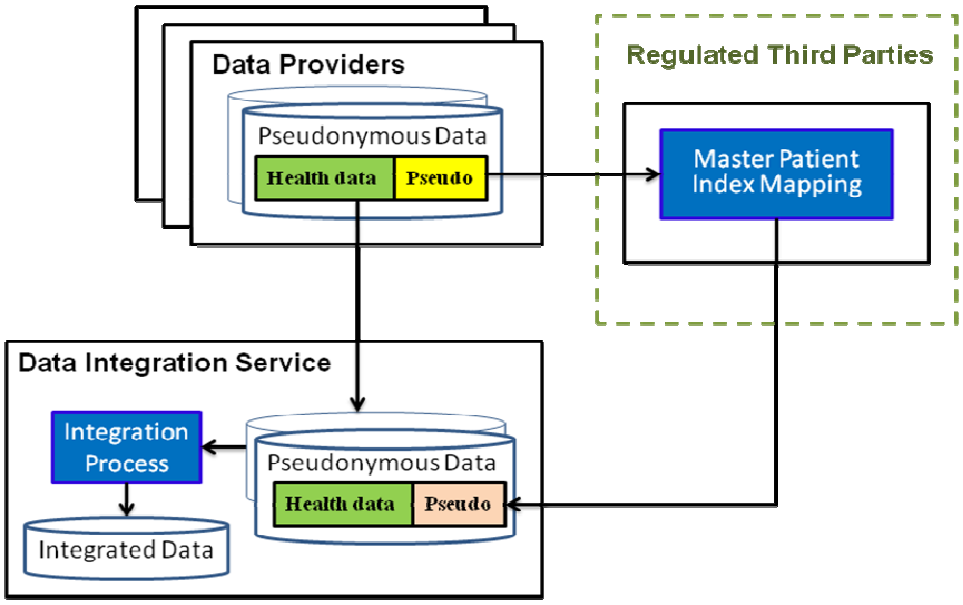

Figure 6-10 Master Patient Index Identity Linking Protocol

Figure 6-11 shows a privacy-preserving query process in Master Patient Index Protocol. The steps of the query are as follows:

1) A consolidated data set request is submitted to the federated data warehouse.
2) The request is analyzed based on the registered data sets and attribute definitions in the Data Set Registry and transformed into sub queries against individual Source Providers that will return partial results that can be integrated into a consolidated data set.

3) Each underlying organization processes its corresponding query. The partial query results include two parts: the Source Provider specific pseudonym and the required health care data.

4) The pseudonyms are sent to the Identity Provider.

5) The other health care data is returned directly to the federated data warehouse.

6) The Identity Provider uses the Master Patient Index to convert each pseudonym into a new pseudonym specific to the consolidated data set and these pseudonyms are returned to the federated data warehouse.

7) The federated data warehouse consolidates all partial query results into a single consolidated data set that is delivered to the user.

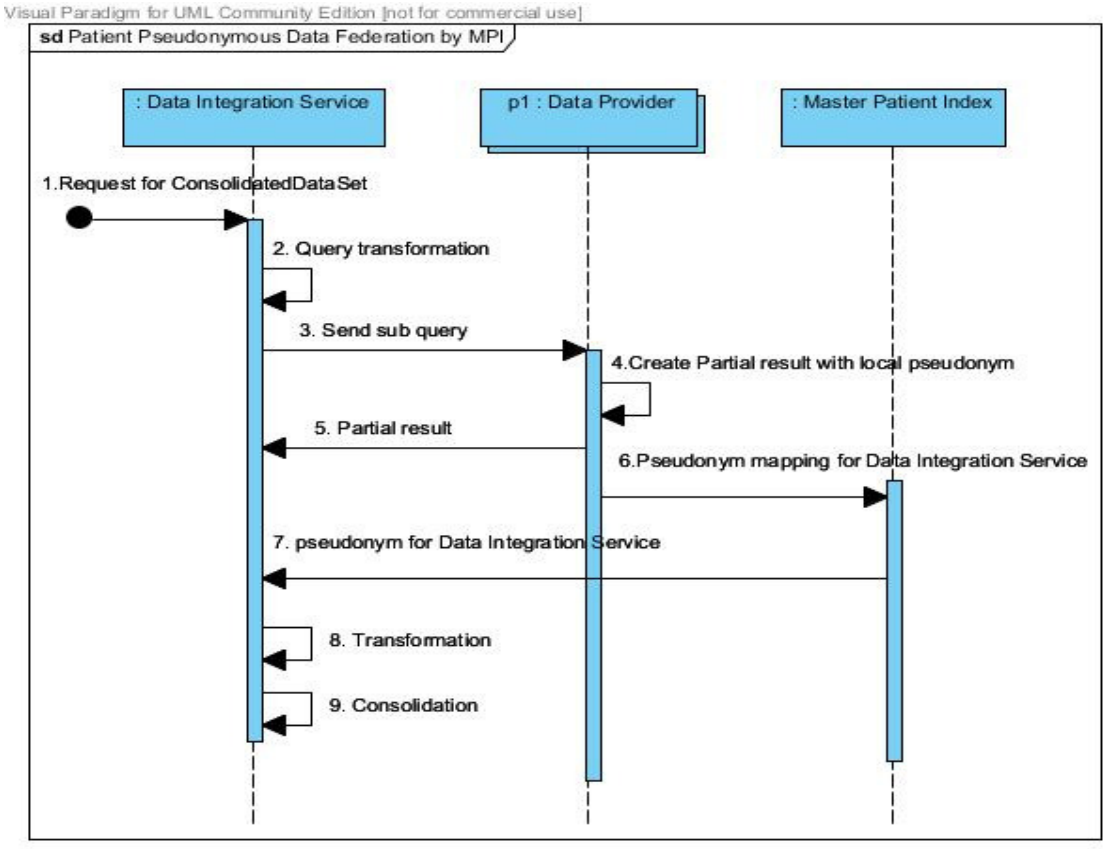

Figure 6-11 Master Patient Index Protocol for Pseudonymous Data Federation

Master Patient Index is good to link identity from different data sources, but security and privacy protection largely depend on the trusted third party and one needs to carefully evaluate the consolidated dataset to reduce the risk of re-identification.

6.4. Anonymized Data Linking

In this section, we describe a set of patterns that can be used to apply different anonymized techniques to hide patient's identity information. Ideally, no identity information should be disclosed apart from the final result. The privacy goals are as follows:

1) **Protection of Patient privacy** Ensure that patient identity will be kept secret from the data integration service and all data providers and ensure that integrated data cannot be re-identified.
2) **Identity Linking.** The identity linking ensures that the data for a patient is linked without revealing their identity, and ensure integrated data cannot be re-identified.

6.4.1 Pattern: Patient Anonymized Fuzzy Hash Linking

Patient anonymized fuzzy hash linking pattern describes how to link and protect sensitive data from multiple data sources by using probabilistic matching on hashed patient identity information.

Context:

Public health surveillance wants to detect unusual trends by collecting, integrating and analyzing the large patient data that is located in different health care organizations. Each organization has its own identity management mechanism.

Problem:

How to collect, link and integrate patient data from multiple distributed organizations while ensuring patient identity is protected?

Forces:

1) Patient data is sensitive.
2) Patient identity information cannot be disclosed.
3) Privacy legislation requires protecting personal data when collecting and using it.
4) There is not a single identifier. Different providers have different id attributes and different formatting.
5) Patient identity is needed to be linked between multiple data providers.
6) Data providers are semi-trusted and are not supposed to collude with each other.

Solution:

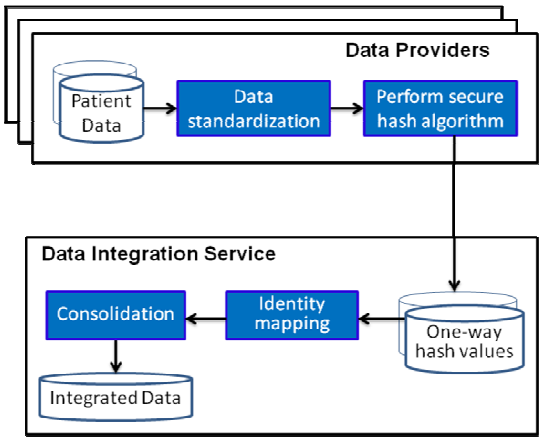

Figure 6-12 Pattern: patient anonymized fuzzy hash linking

Sometimes public health surveillance needs to analyze and detect unusual trends by collecting and linking patient data that is located in different distributed health care organizations. The organizations may have similar patient identity attributes, such as last name, first name, birth date, health card number and etc. But they do not necessarily have the same format or single identifier. In this case, patient anonymized fuzzy hashing linking can be used. It describes how to link and protect sensitive data from multiple data sources by anonymizing patient data.

Patient anonymized fuzzy hashing linking pattern allows linking patient data from multiple data sources while protecting sensitive data by anonymizing data. Patient data should be anonymized before it is sent to the Data Integration Service. As shown as Figure 6-12, each organization performs data anonymization into a one-way hash using same salt value after data standardization, and then sends the hash value to the Data Integration Service which will perform a probabilistic matching algorithm to resolve identities and link patient data. This solution accepts an inconsistent, ambiguous state and does the best to match. An example of such an algorithm is DB2 Anonymous Resolution

(AR) (Swire, 2009). The protocol for patient anonymized data linking for protecting patient privacy using a technique similar to AR is shown as Figure 6-13.

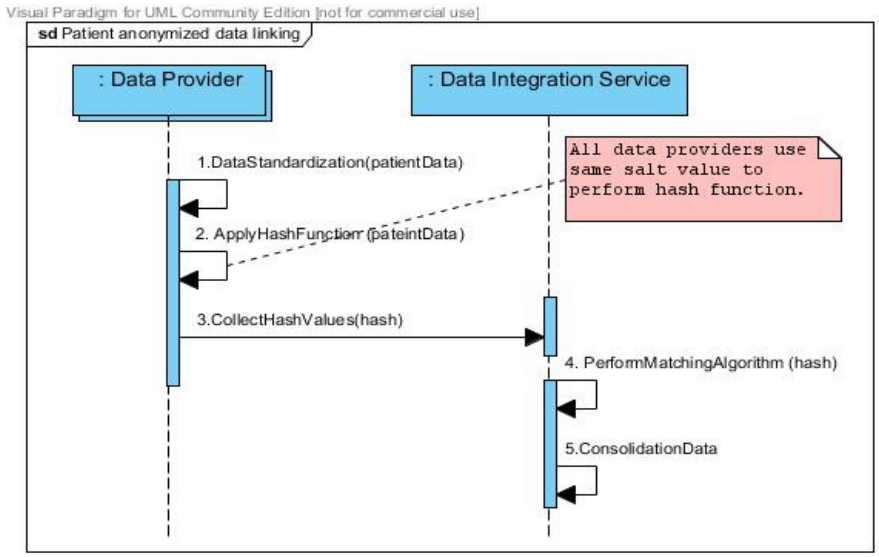

Figure 6-13 Protocol detail: patient anonymized data linking

1) Each Data Provider performs a data standardization process such as data quality, cleanup and normalization process.
2) Each Data Provider performs the secure hash algorithm on each record into a one-way hash. All Data Providers must use same the salt value to anonymize their source data so that when the Data Integration Service site processes the data, it is optimized for record mapping.
3) Each Data Provider sends hash values to Data Integration Service.
4) Data Integration Service executes record mapping algorithm on anonymized records and resolves matching records based on one way hash.
5) Data Integration Service consolidates data from all Data Providers.

Probabilistic matching techniques cannot guarantee 100% accuracy. It depends largely on the completeness and accuracy of the information to be linked and an appropriate combination of matching variables. Therefore data quality and process order is extremely important for this solution.

6.4.2 Pattern: Secure Multi-party Computation Linking

Secure multi-party computation pattern describes how to link and protect sensitive data, and generate the statistics to allow correlation analysis from multiple data sources. This pattern captures the essential elements of the protocol described in section 4.3.

Context:

Public health surveillance wants to conduct the correlation analysis and detect unusual trends by collecting, integrating and analyzing the large patient data that is located in different health care organizations. It requires a set of statistical reports to assist its analysis.

Problem:

How to link and integrate patient data, and generate reports from multiple distributed organizations while ensuring patient identity is protected?

Forces:

3) Patient data is sensitive.
4) Privacy legislation requires protecting personal data when collecting and using it.
5) Patient identity information cannot be disclosed.
6) Linking identity is required.

7) Eliminate risk of re-identification.

8) Providers are not willing to share data based on patient identity.

9) Not using a trusted third party. Semi-trusted third parties are required.

Solution:

Upon receiving a request, each Data Provider executes the query and obtains a dataset, and a patient list. It creates a one way hash for each patient in the dataset based on patient identifier and a private value chosen by each Data Provider, and then sends the generated hash value to Data Integration Service. Each Data Provider agrees upon a unique identifier for the patients. This can be the health care card number, or the combination of name, birth date and gender, or even an identifier generated by a tool. Each Provider has a different private value only known to itself. Consequently they generate different hash values for the same patient. Two semi-trusted third parties are used to coordinate all Data Providers and use a strong hash accumulator. Each of them only possesses partial data, and work with each other to resolve identity linking. It can eliminate the problems with consistent hashing, where the same hash value is broadly used with patients, and it may lead to re-identifications through linking. If the source information is a social assurance number or common demographics, then consistent hashing has high risk of re-identification by exhaustively computing. This pattern, on the other hand, can resist dictionary attacks and organization collusion attacks. In addition, it overcomes the problem of a trusted third party. Figure 6-14 is a pattern for obtaining a count of patients that have attributes in different databases.

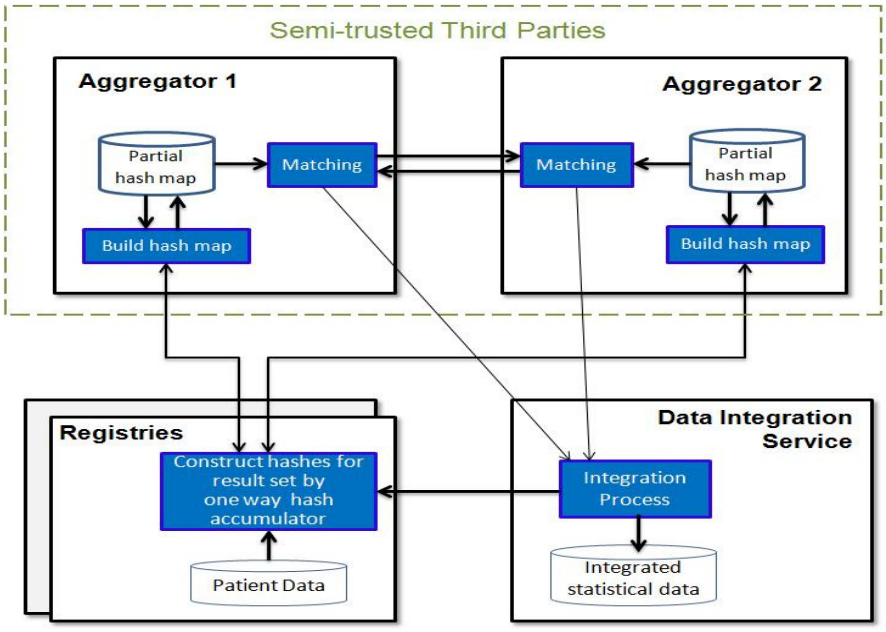

Figure 6-14 Pattern: Secure multi-party computation linking

1) The end user sends a query for obtaining a count of patients that have properties in different databases by Data Integration Service.
2) Data Integration Service transforms it into subqueries for relevant Data Providers.
3) Data Integration Service sends a subquery to relevant Data Providers.
4) Each Provider executes the subquery and constructs ID_1 list for all relevant patients (subquery dataset) by using one-way accumulator.
5) Each Provider sends ID_1 to an Aggregator.
6) This aggregator forwards ID_1 to all other provider to rehash.
7) All other Data Providers rehash ID_1.and send back to the Aggregator.
8) Each Aggregator builds a hash map based on hash and rehash values.
9) Both Aggregators collaboratively compute the matching number and resolve the same identities.
10) Data Integration Service creates final results.

Chapter 7. Case Studies

In this chapter, we will use case studies developed in partnership with health surveillance organizations to illustrate and evaluate how the new protocols and the supporting framework can be used to address real examples of data integration for public health surveillance.

7.1. Secure Computation of Counts for H1N1 Surveillance

The first case study involves the integration and secure computation of counts for surveillance of influenza such as H1N1. To validate our approach, we have done an in-depth case study based on our interactions and experience with the International Society for Disease Surveillance (ISDS, 2010). Much of what is described here was first introduced in (El Emam, Hu & et al, under second review at JAMIA).

7.1.1 Case Study Description

The society wishes to monitor key influenza indicators to determine the severity and spread of influenza, and detect trends in the spread of influenza across North America and eventually the world. The influenza surveillance system needs to collect data from clinics, hospitals and other health practices. To detect meaningful trends, the counts of patients presenting with the particular syndrome, such as gastro-intestinal symptoms (GI) and influenza like illness (ILI) are collected. Individual patient level data is not needed, but age groups and region are helpful to detect and localize epidemics within geographically or demographically defined sub-populations. In our case study, the syndromes are ILI and GI; the age groups are <2, 2-4, 5-17, 18-44, 45-64, 65+ years.

Therefore, 2x6=12 counts should be collected from each data provider every day. The total counts of ILI and GI cases are monitored by patient age groups and geographic groups. All data providers are clustered into groups based on their geographical locations. The counts within one group will be summed up. Typically these counts are coordinated in collaboration with local and provincial (or state) health authorities. Finally the aggregated counts are displayed on a map. The International Society for Disease Surveillance would like to establish national and international integration services to support this.

To implement such a program of public health surveillance, it is useful to follow the nine steps of our framework methodology as described in Section 5-5 The first step is to establish the business relationships and agreements among doctor clinics, hospitals, local health authorities, provincial authorities and national integration service to allow aggregate influenza related data to be collected from different data providers along with proper organizational authorization to use the data in influenza surveillance. The other main steps include selecting a proper privacy preserving data integration protocol, which defines how and when the data will be acquired from providers; preparing data for de-personalization and publishing, then integrating and modeling for reporting; and finally deploying and monitoring the service.

However concerns about privacy and confidentiality often limit the willingness of data custodians to share data for this service. Neither patients nor reporting organizations want to be identified when the surveillance results are public. In particular, the re-identification risk is high when the surveillance results are published in a location based

reporting system. Although aggregate data can be shared if the data is de-identified, the reporting organizations are often unsure of the risk exposure if they disclose patient data on an on-going basis and may be concerned about how data about their organization may reflect on them. Therefore, privacy of patient and data provider should be protected during data collection and data integration. Moreover the appropriate patient consents and/or organizational agreements must be in place. Care must be taken to ensure that the resulting data do not result in situations where patients or data provider become potentially identifiable or be inferred when the results are reported in a map.

7.1.2 Protocol Selection

Currently, the typical approach for influenza surveillance uses the provider anonymized data aggregation pattern (described in Section 6.2.2) where the aggregate totals are collected at a few levels: local, province and national or even international health authorities. Since all counts or totals are clear, the third parties must be trusted which is not secure. Our proposed k-key holder protocol described in Section 4.1 and Section 6.2.3 was developed for this case study to enhance privacy and security to address issues faced by the ISDS. In particular, all of the following forces apply to this case study:

1) Patient data is sensitive.
2) Privacy legislation controls personal data collection.
3) Data providers are concerned about their privacy and risk of reputation management, and may be unwilling to share data.
4) Patient level data is not required.

5) Aggregated counts of patients grouped by some sub-populations are useful.

6) Grouping of data providers is required and useful for reporting, such as location based reporting.

7) Risk of re-identification is high. For example, with location based reporting.

8) Not acceptable to trust a third party. Semi-trusted third parties are required.

7.1.3 Protocol Design

A prototype system was designed and implemented using two aggregators and three key holders as shown in Figure 7-1, but other configurations are possible. The minimum requirement for this protocol is to use one aggregator and two key holders. We choose two aggregators which are fully redundant so that the protocol still work well even one of the aggregators fails. We choose three key holders because the protocol still obtains the correct results even if one of the key holders fails.

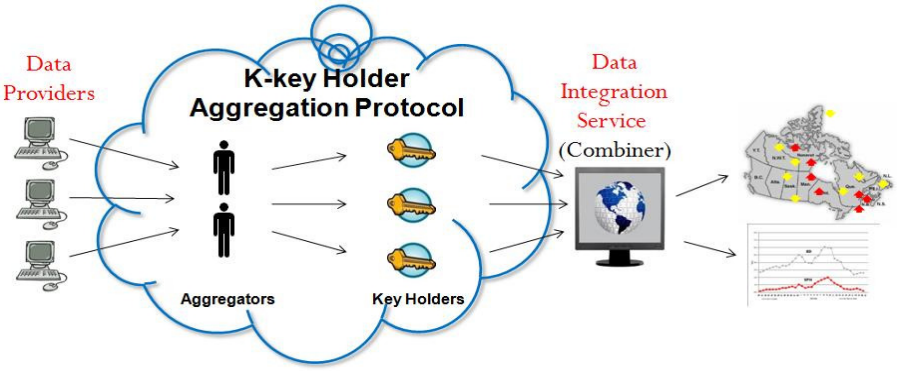

Figure 7-1 k-key holder protocol for secure computation for influenza surveillance

1. Prototype Components

Figure 7-2 shows the kinds of nodes in the system and the kinds of components they hold. Encryption keys of size 512-bit, and a (2, 3)-threshold version of the Paillier algorithm

(Paillier, 1999; Fouque et al, 2000) were used. The communication between the provider, the aggregator service and the data integration service is a standard web service interface using SOAP messages. Basically the system contains the following services:

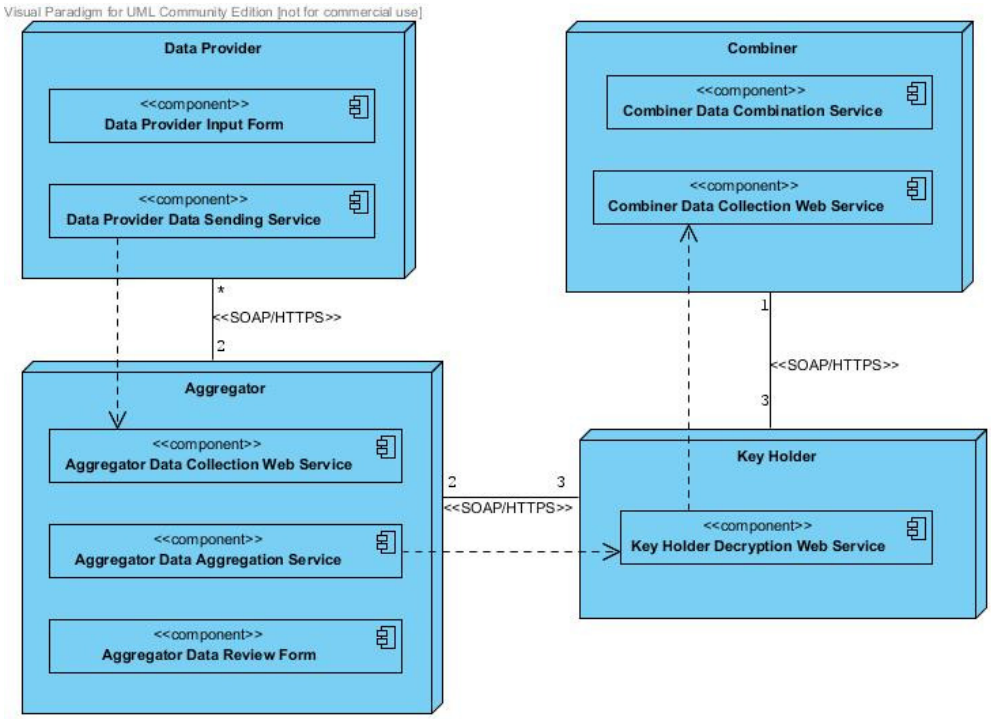

Figure 7-2 Deployment view of secure computation of counts

- In a **data provider**, an interface is implemented to allow the provider to enter the counts and view the history of the counts that have been sent. The screen shots can be seen in Figure 7-3. The counts are persisted in a relational database. A window service runs continuously in the background and is used to encrypt and send the counts to the aggregators as SOAP messages at regular scheduled intervals. The data provider is configured to send its counts to at least two aggregators.

- In an **aggregator**, a web service runs continuously to receive the SOAP messages sent by data providers and to persist the encrypted counts received. A windows service runs continuously in the background and provides an aggregation service that is able to take the encrypted counts and sum them into an encrypted aggregate total for each group using the Pallier algorithm.

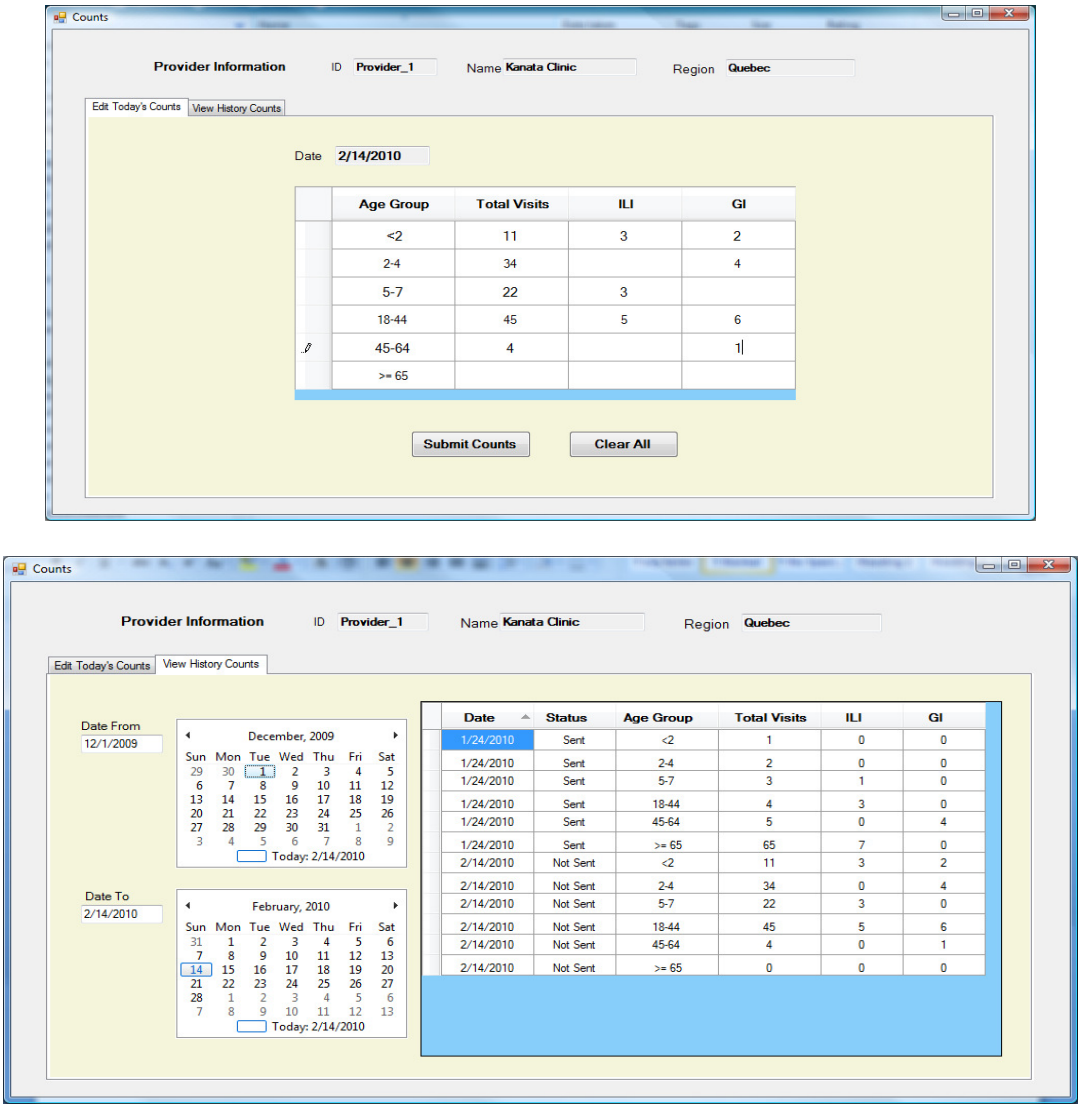

Figure 7-3 Screen shots of a data provider

- In a **key holder**, a web service runs continuously to receive the SOAP messages of encrypted group aggregate totals sent by aggregators. It also partially decrypts them using the Pallier (2,3) threshold partial decryption algorithm and sends these partial decryptions to the data integration service (combiner).

- In **data integration service**, a web service runs continuously to receive the SOAP messages of partial encryptions from the key holders and persists the partial decryptions. A windows service runs continuously in the background and provides a combination service that computes the final counts based on all partial decryptions using the Pallier (2, 3) threshold combination algorithm.

2. Prototype Specifics

To enhance privacy protection, trustiness and robustness of the system, the prototype has the following specifies that were determined in consultation with the ISDS and the public health experts with which we were collaborating:

- To prevent the providers from being re-identified, a minimum of five providers reporting within each region is required in our prototype. If the number of the counts is less than five for a region, we will have a "No Data" results for this region.
- To detect fraud or errors and ensure the trustiness, the verification algorithm in (2,3)-threshold Paillier cryptograph is used. The data integration service will verify the partial decryptions from the key holders using their proof. Only valid decryption results are used for calculating the final results.

- To be robust, all counts from data providers are sent to both aggregators. Two aggregators normally provide the same sum for a region but also it is possible that two aggregators will give different sums if a provider fails to send its data to one aggregator but succeeds with the other aggregator. The data integration service will takes the larger value as the final result.

3. Experimentation

The prototype system was tested for both accuracy and performance using a carefully defined simulation of 3000 data providers and 200 regions for which data is collected and integrated daily over a five week period. A careful analysis of audit trails at each node, showed that all intermediate totals were properly encrypted and that accurate aggregates were obtained in the final data integration. The simulation included simulation of failed messages and downtime for various web services and windows services.

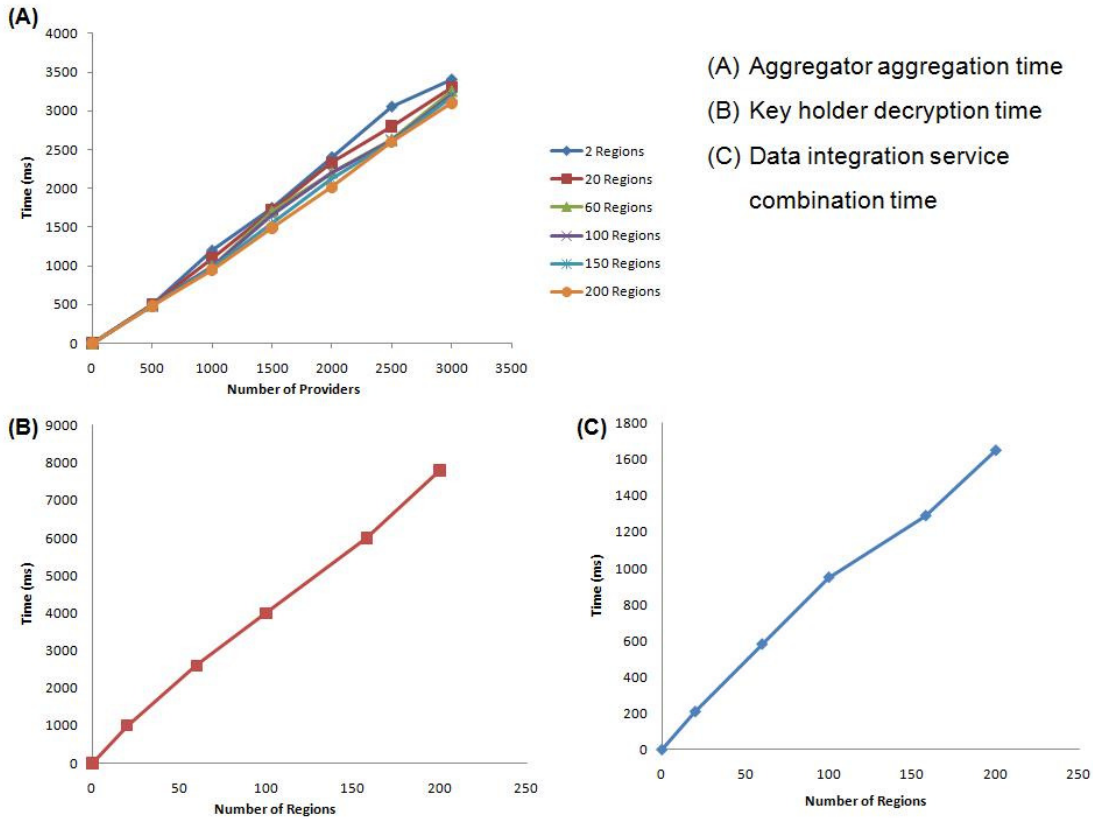

Figure 7-4 Computation time for k-key holder protocol

The performance was measured both in terms of the amount of communication and the time taken to perform the computations. However, since communication time will be variable depending on the nature of the used networks, our empirical evaluation focused on the performance of the computations (encryption, aggregation and decryption) themselves. Our results (see Figure 7-4) showed that the calculation time in aggregators, key holders and the data integration service was linear with respect to the number of data providers and regions. Each data provider encrypts 12 totals once each day when it sends to the aggregators. The aggregators do the aggregation for each region once each day. The average computation (aggregation) time per day for aggregator is around 3 seconds for 3000 providers across 200 regions. The key holders partially decrypt the encrypted

aggregates from the aggregators once per day. The key holder time is under 8 seconds for 200 regions. The data integration service combines the partial decryptions from aggregators once per day, and the data integration service uses less than 2 seconds for the 200 regions. The overall computation time is around 12 seconds.

7.2. Secure Computation for Nationwide HPV Surveillance

In this section, we validate our framework and proposed protocols with an in-depth case study based on our collaboration and experiences in partnership with Health Canada to address human papilloma virus (HPV) surveillance. In this case study, we performed iterative investigation, experimentation, and development of a few candidate protocols that resulted in the protocols presented in sections 4.2 and 4.3.

7.2.1 Case Study Description

The Health Canada researchers we are collaborating with want to conduct HPV surveillance and monitor HPV-related data across Canada. They want to understand the relationship among HPV infection and related outcomes in selected populations. To address the short and long-term objectives of a comprehensive HPV surveillance system, it requires data linked from several population-based data sources including a cancer patient database, cervical screening database, and a heath care services and immunization database. In Canada, these databases are a jurisdictional responsibility. However, the data in these different data sources are not usually linked in a manner that would support integrated HPV infection and related outcomes surveillance. Health Canada would like to integrate data from all the jurisdictions in order to track national trends and generate reports based on statistical analysis such as counts, percentages, chi square, odd ratio and

p-values. In order to achieve this they need to establish a framework (methodology, architecture and protocols) with supporting legal contracts that the jurisdictions will agree to.

We have applied our methodology from section 5.5 and have analyzed current HPV surveillance situations among different jurisdictions within Canada and have separated them into two types, as follows:

Scenario A: Presence of robust and linkable data sources which are used to conduct integrated HPV surveillance within the jurisdiction. However, due to privacy and confidentiality concerns, the linked data is not readily available to inform the national integrated HPV surveillance.

Scenario B: Data on HPV and related outcomes (ex: cancer, pap screening, vaccination) are available within the jurisdiction but these data are managed by separate organizations in separate systems which are not allowed to be linked, making it difficult to inform integrated surveillance initiatives at both the jurisdictional and national levels.

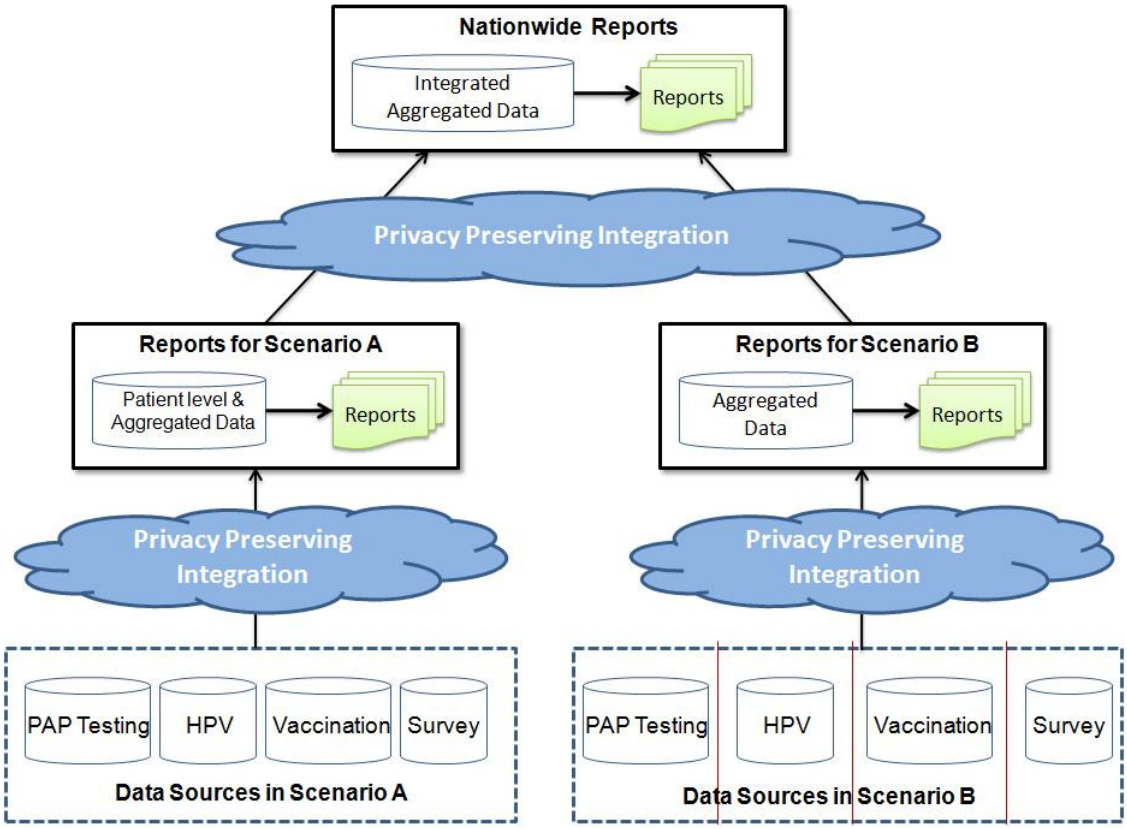

Figure 7-5 Overview of nationwide HPV surveillance

In this case study, we aim to securely integrate data and generate reports for a nationwide HPV surveillance that covers two types of jurisdictions as shown in Figure 7-5. Note that the red dividers for registries in **Scenario B** indicate both separate organizational control and the restrictions on patient-level data. To do so, we are facing the following four challenges:

1) How to build data systems that securely protect patient privacy in jurisdictions for **scenario A** where data records can be linked and integrated in an efficient way based on patient identity.

2) How to integrate data from existing data sources in jurisdictions for **scenario B** where it is not allowed to share patient identifiable information or link patient records within or outside jurisdictions.
3) How to conduct Canada-wide correlation analysis related to integrated HPV and related outcomes surveillance across both types of scenarios.
4) How to reduce re-identification risk when data is disclosed for secondary use.

7.2.2 Protocol Selection

To address the first three issues, we need three different protocols: one protocol for privacy-preserving integration within **Scenario A** to address Issue 1, a second protocol for privacy-preserving integration within **Scenario B** to address Issue 2, and a third protocol for nation-wide privacy-preserving integration to address Issue 3. To address the fourth issue, only aggregated data or statistics are disclosed for second use since aggregate and managed data are at the lowest risk of re-identification (El Emam 2010).

1. Scenario A Privacy-Preserving Integration

Currently, the typical approach for integrating data from the data sources that are willing to share data for disease surveillance just simply replaces person names with pseudonyms, where the same person has the same pseudonym across all data sources (Section 6.4.1). It is not secure and does not resist the attacks like dictionary attack. We developed the federated pseudonymous data linking protocol (Section 4.2 and Section 6.4.2) to address this issue. All of the following forces applied to **Scenario A**:

1) Patient data is sensitive.

2) Privacy legislation requires protecting personal data when collecting and using it.
3) Patient identity information cannot be disclosed.
4) Linking identity is required.
5) Providers are willing to share data based on patient identity.
6) The linking mechanism is reusable to improve efficiency.

2. Scenario B Privacy-Preserving Integration

Currently, it is not possible to integrate data from the registries that are not authorized to share or link patient level data without involving a trusted third party. We developed the secure multi-party computation pattern (Section 5.3 and Section 6.3.2) in order to make it possible to perform the computations needed for HPV surveillance without linking patient records. In particular, all of the following forces applied to **Scenario B**:

1) Patient data is sensitive.
2) Privacy legislation requires protecting personal data when collecting and using it.
3) Patient identity information cannot be disclosed.
4) Linking identity is required.
5) Eliminate risk of re-identification.
6) Providers are not willing to share data based on patient identity.
7) Not using a trusted third party. Semi-trusted third parties are required.

3. Nation-Wide Privacy-Preserving Integration

Currently, the typical approach for integrating data from different jurisdictions is to perform aggregation on clear data (Section 6.2.1 and Section 6.2.2). However, if a

nationwide trend is computed and the provincial values are required to be hidden, a more secure protocol is needed. The k-key holder protocol described in Section 4.1 and Section 6.2.3 combining its supporting framework can be used to enhance privacy, timeliness, and flexibility. In particular, all of the following forces apply to this case study:

1) Privacy legislation controls personal data collection.
2) Data providers are concerned about their privacy and risk of reputation management, and may be unwilling to share data.
3) Patient level data is not required.
4) Aggregated counts of patients grouped by some sub-populations are useful.
5) Grouping of data providers is required and useful for reporting.
6) Risk of re-identification is high. For example, the location based reporting.

7.2.3 Protocol Design

This section describes our design for each of the three protocols.

1. Scenario A Privacy Preserving Integration

There could theoretically be up to four data sources to inform the integrated surveillance of HPV and related outcomes in **Scenario A**:.1) the HPV database which records the HPV types for patients; 2) the PAP testing database which records pap test results for patients; 3) the immunization database which records the vaccination information for patients; and 4) a survey database which records information collected through questionnaires on knowledge, attitude, behaviors related to HPV for the patients. The data sources are trusted and the data can be shared and linked but privacy legislation still

requires the organizations to put in place appropriate safeguards and minimize the risk. Protecting patient privacy needs to limit the use of identity information to be only on a need to know basis. A Master Patient Index (MPI) ensures that identifying information is separated from sensitive health care data; pseudonyms are used to link data while still masking patient identity. The organization needs to be trusted because it is possible to re-identify using the MPI, but the chances for disclosure are minimized and need to be more carefully controlled. When designing such a surveillance system for **Scenario A**, we focus on the following requirements: separating patient identity from other data; protecting patient identity when linking; patient level data consolidation and linking are strict on a need to know basis; and only publish aggregated data.

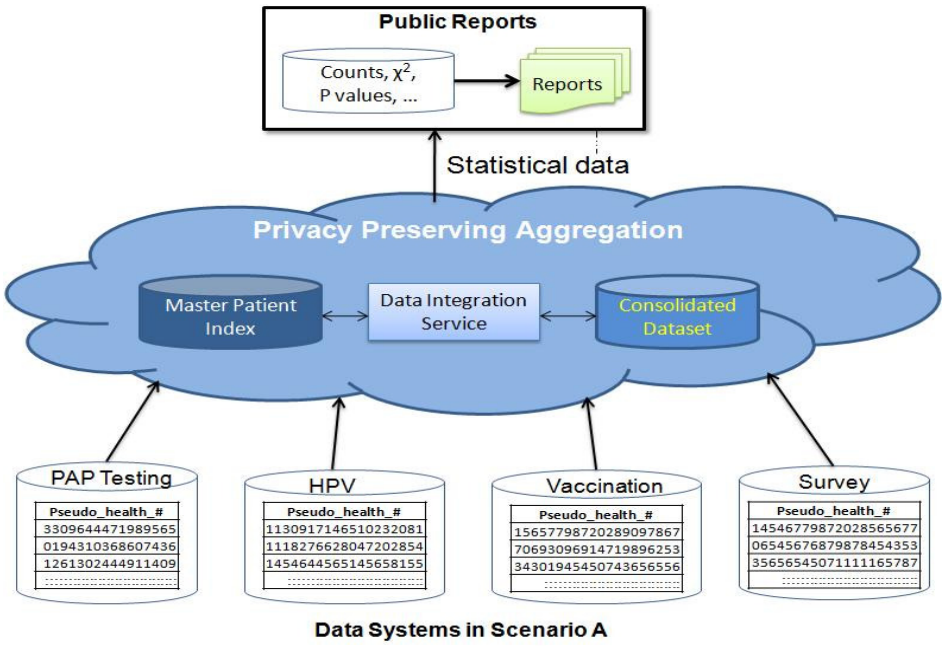

Figure 7-6 Scenario A Privacy Preserving Integration

Figure 7-6 shows overview of the architecture and protocol. The detailed protocol is described in Section 5.2. Besides the four data sources, we have two additional entities,

a Data Integration Service and an Identity Provider as described in Figure 5-2. The Data Integration Service is used for integrating data across registries and creating consolidated datasets by coordinating Identity Provider. Identity Provider conducts a federated pseudonymous identity management among data sources. Figure 7-7 shows how to create a consolidated data set from data sources using the Master Patient Index and Data Integration Service. It improves the protection of patient privacy because those working with registries are not able to know identity, and those working with identity are not able to know health data. But this is only true if the Identity Provider is a separate trusted organization (i.e. the Identity Provider will not collude with the data source organization or the data integration service). Also, it becomes easier to re-identify the more data is linked together, so the Data Integration Service must be trusted and as described in our methodology (refer to Section 5-5) there needs to be a strict process in place for evaluating the potential for re-identification of any data sets created.

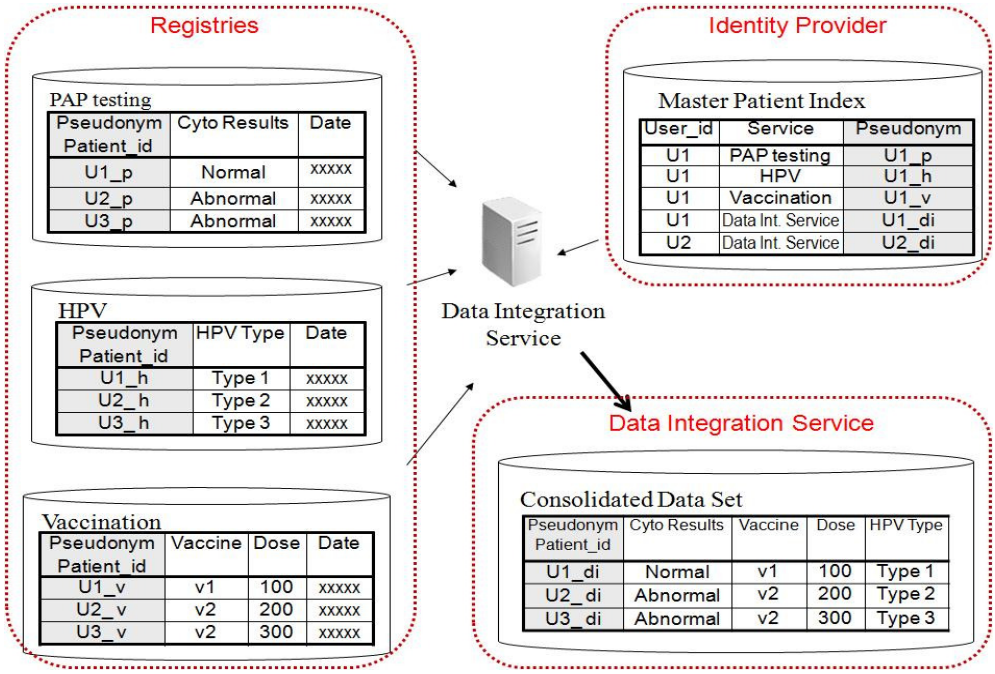

Figure 7-7 Creating consolidated data set by Master Patient Index

2. Scenario B Privacy Preserving Integration

In **scenario B**, similar to **scenario A**, there will be four data sources to inform integrated HPV and related outcomes surveillance: 1) the HPV database which records the HPV types for patients; 2) the PAP testing database which records pap test results for patients; 3) the immunization database which records the vaccination information for patients; and 4) a survey database which records information collected through questionnaires on knowledge, attitude, behaviors related to HPV for the patients. Although aggregate data from each database can be used to inform the integrated HPV surveillance, data linkages between the data sources based on patient identity are not allowed. When designing such a surveillance system for **Scenario B**, we focus on the following requirements: patient data is anonymized (data is encrypted) before sending out for aggregation so that no personal information is disclosed; only aggregated data or statistics are created for disease surveillance so that re-identification risk is reduced to a minimum.

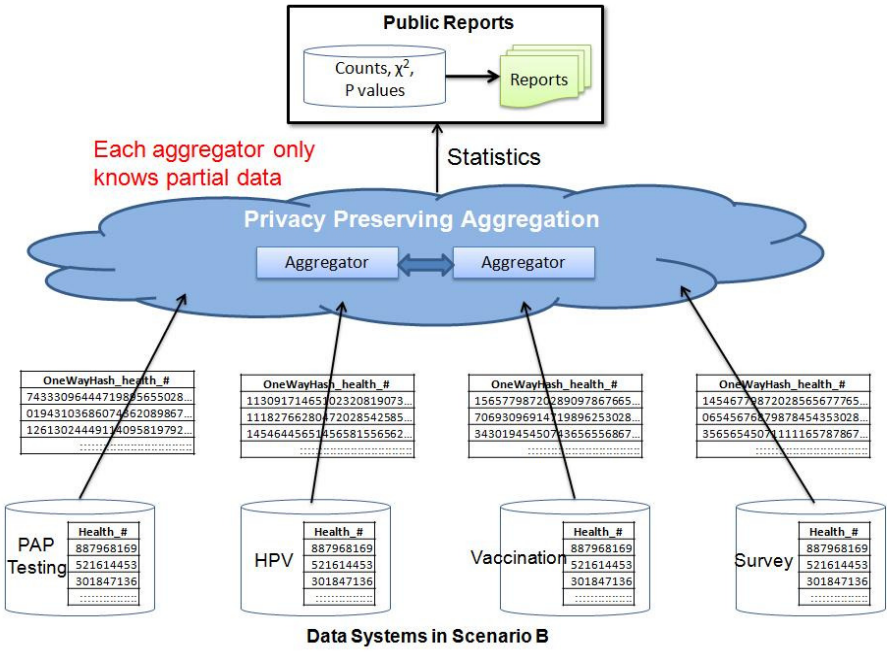

Figure 7-8 Scenario B Privacy Preserving Integration

Figure 7-8 shows overview of the architecture and protocol. The detailed protocol is described in Section 4.3. As shown in Figure 7-8, besides the four data sources, in our design there are two aggregators. When answering the query, each data source sends its results to an aggregator that is randomly chosen. Therefore each Aggregator only knows a partial of the whole result. Accurate aggregates, that would normally require linking of patient-level data, can be computed across all four registries through multi-party secure computation. In principal, any query that provides aggregate results can be computed, although each such query requires its own special setup and configuration. Careful attention must be paid to population sizes when looking at the individual aggregate values that result from a query. Special processes must be put in place to handle the "small cell" problem where there is a risk of re-identification when aggregate values are small. For example, if the adversary can determine that there is only one aboriginal individual who does not have a HPV in the population, then the re-identification risk is high. Figure 7-9

shows a more detailed architecture, where the aggregators and the data integration service are the same entities mentioned in the architecture as Figure 5-2 and the data sources are the data providers from Figure 5-2.

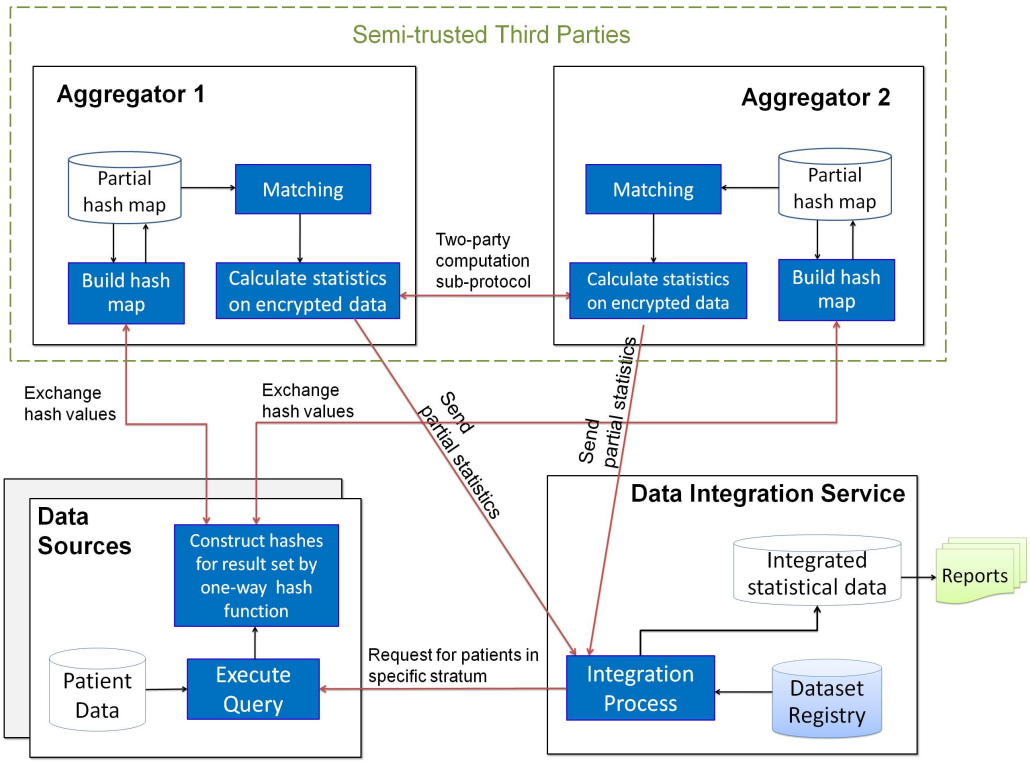

Figure 7-9 Secure multiparty computation components

- A **Data Source** is able to execute the query and get the datasets. It is also able to construct the hashes for the datasets by one-way hash function. In addition, it can re-hash the values sent by aggregators.
- An **Aggregator** can build a hash map by communicating with Registries and performing the mapping functions. It also collaborates with the other aggregator to calculate the counts or statistics using a two-party secure computation sub-protocol.
- The **Data Integration Service** is responsible to send requests to Registries. It also maintains a dataset registry that manages datasets for contingency tables. In

addition, it is able to compute the final results based on the partial values from two Aggregators.

3. Nationwide Privacy Preserving Integration

It is relatively straight forward to aggregate totals from jurisdictions in Scenario A and jurisdictions in Scenario B for nationwide surveillance following a similar approach to our first case study in 7.1. Since the data collected from jurisdictions is already aggregated, the risk to patient privacy is minimized. But, it is important for all jurisdictions to be able to standardize and agree upon a common view or set of views for querying data across organizations and integrating the resulting data sets. In some cases, a nationwide report can be generated based on plain aggregates and/or statistics that are submitted from all jurisdictions. However, if a nationwide trend is computed and the jurisdictional values are required to be hidden, a more secure aggregation protocol (see the protocol describe in Section 5-1) is needed to protect the confidentiality of jurisdictions. In this case, the data from jurisdictions are encrypted before they are sent for aggregation so that no party will know the exact value from a particular jurisdiction.

4. Summary

In this case study, we have applied our framework and methodology to analyze different situations in two scenarios and identify four major requirements we need to solve. Based on the trust levels of stakeholders, potential of re-identification risk and identity protection requirements, we created three different protocols, Master Patient Index linking protocol (refer to Section 4.2 and Section 6.3.2) and multi-party secure computation protocol (refer to Section 4.3 and Section 6.4.2) and k-key hold provider

anonymized data aggregation protocol (refer to Section 4.1 and Section 6.2.3) for privacy preserving data integration in two different scenarios. Then the aggregate data from all jurisdictions are integrated to generate nationwide reports.

This case study illustrates that the pattern classification is useful for understanding and selecting proper protocols to meet the specific requirements of each scenario. A proper privacy-preserving protocol relies on a proper identity protection, either anonymized or pseudonymous. In addition, a reference architecture is useful and necessary for privacy protection. The fact that third parties are not required to be trusted but only need to be semi-trusted is important. For example, two semi-trusted third parties are useful for maintaining the confidentiality of the results. The dataset registry within the data integration service is also important for querying and managing contingency table and datasets. As well, the methodology described in Section 5.5 guides the data integration process that emphasizes privacy protection and minimizes the risk of re-identification.

Chapter 8. Framework Evaluation

In this section, we will evaluate our framework based on our case studies which were done in collaboration with public health organizations and their experts in data integration and privacy to see how well it has addressed the criteria we identified and defined in Section 3.3. We will also compare our framework with existing solutions in terms of the details of our proposed architecture, methodology, and our proposed new data integration protocols.

8.1. Evaluation of Overall Framework

Table 8-1 evaluates how well our proposed framework addressed the evaluation criteria in Section 3.3 when applied to our case studies. The first column corresponds to the criteria in Section 3.3. The architecture column identifies which components of our proposed architecture from Section 5.4 address the criteria. The methodology column identifies which steps of our proposed methodology from Section 5.5 address the criteria. Finally, the last column identifies which or our three proposed new protocols from chapter 4 address or do not address the criteria.

Table 8-1 Components in proposed framework to meet the criteria

Criteria	Privacy-Preserving Data Integration Framework		
	Architecture	**Methodology**	**Proposed Protocols**
1-a. Support distributed data sources.	Data Integration Service Dataset Registry	4. Data Publishing 5. Data Integration 9. Monitoring	All three protocols
1-b. Support near	Data Integration	5. Data Integration	All three protocols

real time data integration.	Service		
1-c. Enable data publishing and data reporting.	Data Integration Service	6. Modeling 8. Deployment	All three protocols
2-a. Ensure that the patient identity is kept secret.	Identity Provider Key holder committee Aggregators	3. Data Preparation 4. Data Publishing 5. Data Integration	Key-key Holder Provider Anonymized Aggregation protocol; Secure Multi-party computation Linking protocol; (Master Patient Index Linking protocol only IF Trust Identity Provider)
2-b/4-b. Ensure that integrated data cannot be re-identified.	Data Integration Service Dataset Registry	6. Modeling 9. Monitoring	All three protocols need to manage small cell problem. Master Patient Index Linking protocol need to trust Identity Provider
2-c. Ensure that patient consents and/or organizational agreement are in place.	Identity Provider Or Trusted Data Providers	3. Data Preparation 5. Data Integration 9. Monitoring	Only Master Patient Index Linking protocol via Identity Provider.
3-a. Ensure linking identity without revealing identity info.	Identity Provider Master Patient Index	5. Data Integration	All three protocols
4-a. Ensure that the data provider identity is kept secret.	Key Holders Aggregators	4. Data publishing 5. Data integration	Only Key-key Holder Provider Anonymized Aggregation protocol
5-a. Ensure integrity of data.	Key Holders Aggregators Data Integration Service	3. Data Preparation 4. Data Publishing 5. Data Integration 6. Modeling	Key-key Holder Provider Anonymized Aggregation protocol; Secure Multi-party computation Linking protocol; (data is not encrypted in Master Patient Index Linking protocol)
5-b. Authenticate sensitive data sources.	Identity Provider Key Holders Aggregators Data Integration Service	5. Data integration	All three protocols

	Dataset Registry		
5-c. Prevent adversary attacks such as organization collusion attack.	Key holders	5. Data Integration 9. Monitoring	Key-key Holder Provider Anonymized Aggregation protocol; Secure Multi-party computation Linking protocol; (Master Patient Index Linking protocol must trust Identity Provider)
5-d. Control Access to sensitive data.	Data Integration Service Dataset Registry	6. Modeling 9. Monitoring	All three protocols

In the table, we can see that our framework provides reasonably complete overall coverage of the evaluation criteria in general, but for specific situations one needs to understand carefully the requirements and forces at work in order to select the appropriate protocol, understand which architectural components are involved and the role they play, and ensure that appropriate steps are taken according to the methodology in order to fill in or address any gaps.

In general, the aggregate and anonymized protocols are more secure in protecting privacy and guarding against collusion, whereas the pseudonymous protocol requires the identity provider to be trusted. It does however have the advantage of providing centralized control and coordination for managing patient consents or organizational agreement.

The three new steps (Data Publishing, Data Integration and Monitoring) we introduced when creating our methodology based on CRISP-DM are the most critical for ensuring coverage of the criteria, although the modifications to data preparation,

modeling and deployment are also relevant. The most significant aspect is the management of the possibility for re-identification.

In terms of architecture, the data integration service and data set registry are most significant for enabling data integration in a B2B environment, while the key holders and aggregators are the most important for security and ensuring the integrity of data.

8.2. Architecture Comparison

Table 8-2 compares the proposed architecture with some of the architecture approaches discussed in chapter 2. Basically a data warehouse (Section 2.2.2) enables historical data integration from different sources but does not takes account of privacy protection; Liberty Alliance (Section 2.3.1) is strong in security and privacy for B2B networks but data integration really only happens one record or one request at a time which is limiting for public health surveillance; Software as a Service (SaaS) (Section 2.2.3) is designed for data sharing among different platforms but data integration and privacy protection need to be designed on an individual process basis. In our framework, a centralized Identity Provider and Master Patient Index ensure trusted centralized third party control of consent, ID consistency, and identity linkage, as it is done in the Liberty Alliance approach. Identity information is protected solely by the Identity Provider, and privacy is safeguarded since the Identity Provider has no access to health care data. Further, the use of a Data Set Registry to define a common data view provides federated data warehouse solution and providers consistency by allowing data sets and attributes to be registered, but is not static since new data sources and attributes can be added flexibly. Standardization is mandatory only within the processing of a single data set not across all

Source Provider databases. In addition, the Data Set Registry provides a centralized mechanism for evaluation of consolidated data sets to ensure appropriate authorization and/or consent is in place. Moreover semi-trusted Third Parties such as Aggregator and Key Holder Committee play an important role in protecting privacy. Aggregator(s) performs aggregation on encrypted data and Key Holder Committee jointly decrypts an encrypted data. Multiple Aggregators and multiple Key Holders ensure confidentiality of true data.

Table 8-2 Architecture Comparison

Criteria	Proposed Architecture	Data Warehouse	Liberty Alliance	Software as a Service
1-a. Support distributed data sources	Data Integration Service Dataset Registry;	Static common data model; ETL, history data	Only one record at a time	Individual process
2. Ensure patient privacy	Identity provider,	No	Identity provider	No
3. Identity Linking	Master Patient Index in Identity Provider	Difficult if no common identifier.	Master Client Index in Identity Provider	No.
4. Ensure provider privacy	Trusted/semi-trusted third parties	No	No	No
5. Data Protection	De-identification; Multi-party secure computation third parties; Access control in Dataset Registry	N/A	Based on Circle of Trust	N/A

8.3. Methodology Comparison

Our framework has advantages in terms of platform, availability, data used, authorization, privacy protection, data sharing, data integration, modeling techniques, monitoring and security. It supports B2B networks over the Internet on a continuous, ubiquitous basis so any organization is potentially a source of data, public health surveillance collaborator and/or consumer of data. Whereas the traditional model is Intranet or Standalone oriented, planned and deployed by a single organization. The protection of sensitive data in our framework is based on patient consent, business agreement, identity protection techniques, and real-time monitoring of regulatory compliance. Data sharing and integration leverage modern Internet technologies. Table 8-3 shows above essential elements that are supported by our framework in comparison with the traditional approach to data mining as defined by CRISP-DM (Section 2.2.1) and Health Research Ethics Board approval process (Section 2.1.1).

Table 8-3 Methodology Comparison

	Privacy Preserving Methodology for Public Health Surveillance (Section 5.5)	**CRISP-DM Process Model (Section 2.2.1)**	**Health Research Approval Process (Section 2.1.1)**
Platform	B2B network, Internet	Intranet , Standalone	Case by case data feed request
Availability	Continuous	Planned	Planned for specific research period
Data	Multiple sources on the Internet from different organizations	Single organization	Single organization, but can combine from several on an ad hoc basis.
Authorization	Permission-based, Automatic	Commission-based	Commission-based
Trust model	Tailorable as needed by appropriate protocol selection.	Single organization	Ethical/privacy review board from EACH data

	Flexible support in architecture through third parties: key holders, identity provider etc. Data Integration Service records agreements as needed.		provider.
Privacy protection	Business agreement, Patient consent form, Regulatory compliance, A spectrum of privacy enhanced techniques	Single organization privacy policies. Patient consent.	Review board oversight of approved protocols
Data sharing	Common data view, Web Services, Data feeds	Enterprise data model	Case by case request for specific questions
Data integration	Virtual data warehouse, Mashups Dynamical and flexible configuration Efficient, lower cost development	ETL Static integration per project. High cost development	Case by case request for specific questions
Modeling techniques	Statistics, Data mining algorithms, OLAP model, Text mining, Streaming Event Data	Statistics, Data mining algorithms OLAP model	Statistics, Data mining algorithms OLAP mode, Text mining
Monitoring	Real-time, dynamic, continuous	N/A	Review board oversight
Security	A spectrum of secure protocols	Single Organization security.	Honor system and approved protocols

8.4. Comparison of Data Aggregation Protocols

In chapter 4.1, we illustrated a new protocol for provider anonymized aggregation. We have also described other protocols in chapter 6.2 that addressed the data aggregation pattern for privacy preserving data integration. In Table 8-4, we compare these protocols with respect to the aspects and criteria identified in Section 3.3 based on our case study described in Section 7.1.

Table 8-4 Comparison of Anonymized Aggregation Protocols

	K-Key Holder Provider Anonymized Aggregation (Section 4.1 & 6.2.3)	Provider Anonymized Aggregation (Section 6.2.2)	Patient Anonymized Data aggregation (Section 6.2.1)
1. Data integration	Suitable for counts, rates	Suitable for counts, rates	Suitable for counts, rates
2. Patient privacy protection	Anonymization by patient aggregation and encryption	Anonymization by patient aggregation	Anonymization by patient aggregation
3. Identity linkage	N/A	N/A	N/A
4. Provider privacy protection	Anonymization by provider aggregation and encryption using semi-trusted Third Party	Anonymization by provider aggregation using Trusted Third Party	N/A
5. Data protection	No patient level data; Prevent re-identification by collecting aggregated data and publishing provider aggregated data; Prevent single adversarial party and collude attack between two organizations by using multiple key holders; Authenticate data source using digital signature	No patient level data; Prevent re-identification by collecting patient aggregated data and publishing provider aggregated data	No patient level data; Prevent re-identification by collecting patient aggregated data.

8.5. Comparison of Pseudonymous Data Integration Protocols

In chapter 4.2, we illustrated a new protocol for federated pseudonymous linking. We have also described other protocols in chapter 6.3 that address the Pseudonymous data integration pattern. In Section 6.3.1, Pseudonymous Data Federation uses one-way hash similar technique determinedly and securely links records. Master Patient Index Data Linking protocol in Section 6.3.2 and leverages centralized Identity Provider and Master Patient Index ensure trusted centralized third party control of consent, ID consistency, and identity linkage. Identity information is protected solely by the Identity Provider, and privacy is safeguarded since the Identity Provider has no access to health care data. In

Table 8-5, we compare these two approaches with respect to the criteria identified in Section 3.3 based on our case study described in Section 7.2.

Table 8-5 Comparison of Pseudonymous Data Integration Protocols

	Master Patient Index Linking (Section 4.2 & 6.3.2)	**Pseudonymous Data Federation (Section 6.3.1)**
1. Data integration	Enable patient-level data integration	Enable patient-level data integration
2. Patient privacy protection	Master Patient Index creates different pseudonyms for same patient in different data providers; Not need to trust data providers; Need to fully trust Identity Provider	Same pseudonym for same patient in different data providers; Need to fully trust data providers;
3. Identity linkage	Pseudonymous; Master Patient Index, Deterministic matching	Pseudonymous; Deterministic matching, Need common identifier
4. Provider privacy protection	N/A	N/A
5. Data protection	Separate identity information from health data; Prevent single adversarial party; Prevent dictionary and collusion attack among data providers.	Mask identity; Not prevent dictionary attack and collusion attack among data providers.

8.6. Comparison of Anonymized Data Linking Protocols

In chapter 4.3, we illustrated a new protocol for anonymized data linking. We have also described other protocols in Section 6.4 that address the anonymized identity linking data integration pattern. In Section 6.4.1, Patient Anonymized Fuzzy Hash Linking protocol resolves ambiguous, anonymized data to identify potential identity matches. Multi-party Secure Computation Linking protocol in Section 6.4.2 adapts a strong hashing function to combine patient and provider to produce pseudonyms for the patient without risk of re-identification. In Table 8-6, we compare these approaches with respect to the criteria identified in Section 3.3 and our case study in Section 7.2.

Table 8-6 Comparison of Anonymized Data Linking Protocols

	Multi-party Secure Computation Linking (Section 4.3 & 2.2.1)	**Patient Anonymized Fuzzy Hash Linking (Section 6.3.2)**
1. Data integration	Enable anonymous data integration	Enable anonymous data integration
2. Patient privacy protection	Inconsistent hash to hide identity from One-way accumulators	Anonymization by one-way hash
3. Identity linkage	Linking individual patient if common identifier is predefined; Anonymous; Hash, re-hash; Multi-party computation using their private values; Need predefined identifier	Linking individual patient record without common identifier; Anonymous, probabilistic matching
4. Provider privacy protection	N/A	N/A
5. Data protection	Use different hashes to same identifier in for different data providers; Prevent re-identification by publishing aggregates and statistics; Authenticate data source by membership testing using one-way accumulator;	Use same hash to same identifier for different data providers.

In addition, similar to Master Patient Index Data Linking (Section 6.3.2), Multi-party Secure Computation Linking protocol (Section 6.4.2) uses different pseudonyms (hashed value) for the same patient; but this hashed value is used only one time while the pseudonyms in MPI is permanent. Master Patient Index Data Linking protocol needs a trusted third party and uses MPI mapping to identify the same patient while multi-party secure computation linking protocol only need semi-third party and it uses multiple calculations to identify the same patient.

Chapter 9. Conclusions and Future Work

This thesis is the result of several years of research in public health surveillance and technologies related to data integration and privacy that included meaningful collaboration with two major public health surveillance organizations where we were privileged to work with a number of experts in public health surveillance and privacy.

The insight we have acquired into the nature of the challenges facing public health surveillance is reflected in the example scenarios we introduced in chapter 3 to identify, at a high level, three main types of data integration problems. This was used to drive our gap analysis of existing approaches and technologies and identify the main criteria which our thesis work addresses. It also drove the development of our pattern classification for privacy-preserving data integration protocols which was presented in chapter 6 to catalog and distinguish the various protocols we encountered in our research as well as the three new protocols we proposed in chapter 4 (one for each main type of data integration problem).

Our work with public health organizations developing those three protocols, as described in our case studies in chapter 7, highlighted the importance of having a framework to guide the implementation. The major opportunity, and the major challenge, facing privacy preserving data integration for public health surveillance, is the pervasiveness of Internet technologies for collecting, sharing, processing, analyzing and distributing data. In chapter 6, we presented an architecture and methodology that enabled continuous privacy-preserving data integration across a B2B network for public

health surveillance. Our proposed methodology updates and extends the CRISP-DM methodology for data mining that has been adopted as a standard by the European Union while our architecture augments the concept of a Circle of Trust with key architectural components to address issues of trust and risk when doing data integration in a B2B network.

In conclusion, we list here again the major contributions of our thesis from chapter 1, and highlight for each the significance of the contribution to research for privacy-preserving data integration in public health surveillance:

- **The development of three new protocols for privacy-preserving data integration for public surveillance**

Three new protocols are more secure than those that are commonly used in current health surveillances. The k-key holder provider anonymized aggregation protocol and multi-party secure computation protocol prevent single adversarial third party attack by using semi-trusted party, which overcome the requirement of all data providers having to fully trust a third party. In addition, both protocols aim to collect and publish aggregates, which reduce re-identification risk to a minimum. In particular, the k-key holder protocol is the "first" protocol in the literature to protect provider identity. The federated pseudonymous linking protocol separates identity information from health data and can prevent collusion attack among data providers.

- **A set of privacy preserving data integration patterns systematically classified based on essential principles for privacy protected data integration.**

Privacy preserving data integration patterns fill a gap in the security pattern literatures. It plays an important role for public health organizations to select a proper data integration solution to meet different privacy and trust requirements in different scenarios.

- **An architecture for continuous privacy-preserving data integration within B2B health care networks**.

The common architecture sets up a single infrastructure that allows different privacy preserving data integration protocols to be implemented. Trusted and semi-trusted third parties are used for enhancing privacy protection. And dataset Registry eanbles inter-organizational data management and access control in a B2B health care network.

- **A methodology for organizing and managing continuous privacy-preserving health surveillance in a B2B network**.

The methodology provides guidance for public health organizations to conduct data integration projects while preserving privacy. It addresses privacy requirement in B2B health surveillance which are not covered in the traditional model (CRISP-DM). It also enables near real time and continuous data integration which are not addressed in the current approach (health research approval process).

Future Work

Despite the value of this research, there are a few limitations that should be addressed in future work.

First, our framework assumes the underlying Internet infrastructure used in our solution is secure. We assume the network is secure, the hardware is secure, and the web server is secure as we focus on issues directly related to privacy when doing data integration. A complete threat risk assessment should be done which should include a comprehensive threat model for our proposed architecture and methodology.

Second, our framework does not cover all aspects of the methodology and architecture for privacy-preserving data integration in enough detail. For example, although patient consents or organizational agreement is an essential aspect of privacy-preserving data integration, our framework does not have a detailed process on how to obtain and manage them.

Third, since our research adopts a design-oriented research methodology, that focuses on demonstrating the utility of our framework for addressing specific identified gaps. A more comprehensive set of trials and empirical evaluation is needed to evaluate our framework to understand the full implications of its use. In addition, future work beyond this thesis might more systematically survey experts and analyze requirements across a broader spectrum of public health organizations than the few we have been able to interact with.

Finally, based on the research results drawing from this thesis, the health care system needs a more systematic and standardized approach to manage identity as well as access to data across organizations for public health surveillance. Because of legal, political, and organization implications, different government jurisdictions will take different approaches. A more comprehensive research is required to establish implementation standards and mechanisms to address legal, political, organizational, and technical challenges.

In addition to addressing the above limitations, there is additional future work that can be done to continue the development and refinement of our framework. The architecture and methodology in our proposed framework require further study in a number of areas. Privacy-preserving data sharing, privacy-preserving data integration and monitoring are new phases introduced to privacy-preserving knowledge discovery and public health surveillance. Each phase itself is a complex process that can be further investigated. The protocol, algorithms, standards, and tasks for each phase are all good opportunities for future research. In particular, how to handle incomplete data and how to link legacy data need more investigation. Moreover there are still many organizational challenges to be faced in setting up and leveraging the existing technologies we have described here, in addition to the new ones we have proposed. Many of them are not yet commonly used.

REFERENCES

Aboulnaga, A., & El Gebaly, K. (2007). µBE: automatic source selection and schema mediation for Internet scale data integration. *In Proceedings of IEEE International Conference on Data Engineering* (pp.186-195). Istanbul, Turkey: IEEE.

Adragna, L. (1998). Implementing the enterprise master patient index. *Journal of AHIMA, 69*(9), 46-52.

Agrawal, D. and Aggarwal, C. C. (2001). On the design and quantification of privacy preserving data mining algorithms. *In Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* (pp. 247–255). Santa Barbara, California, USA: ACM.

Agrawal, R., Evfimievski, A., Srikant, R. (2003). Information sharing across private databases. *In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, San Diego, CA, 2003, pp.86-97.

Agrawal, R., Asonov, D., Srikant, R. (2004). Enabling sovereign information sharing using web services. *In Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, 2004, pp. 873-877.

Ainsworth, J., Crowther, P., Buchan, I. (2009). Federating Health Information Systems to Enable Population Level Research. *In 22nd IEEE International Symposium on Computer-Based Medical Systems*, 2009. CBMS 2009.

Baron, S., Spiliopoulou, M., & Günther, O. (2003). Efficient monitoring of patterns in data mining environments. *Advances in Databases and Information Systems* (pp: 253-265). Berlin / Heidelberg: Springer.

Benaloh J. C. (1987). Secret sharing homomorphisms: Keeping shares of a secret secret. *In A. Odlyzko, editor, Advances in Cnjptology, Proc. of Crypto '86.* Lecture Notes in Computer Science 263), pp. 251-260. Springer-Verlag, 1987. California, U.S.A., August 11-15.

Benaloh J. and de Mare M. (1994). One-way accumulators: a decentralized alternative to digital signatures. *In Proceedings of Advances in Cryptology - EUROCRYPT '93*, Lecture Notes in Computer Science, v 765, pages 274-285, Lofthus, Norway, 1994.

Berman, J. J. (2004). Zero-check: A zero-knowledge protocol for reconciling patient identities across institutions. *Archives of Pathology & Laboratory Medicine.* 128 (2004), 344-346.

van Blarkom, G.W., Borking, J.J., Olk, J.G.E. (2003). PET. *Handbook of Privacy and Privacy-Enhancing Technologies*. ISBN 90-74087-33-7.
http://www.andrewpatrick.ca/pisa/handbook/handbook.html, last retrieved November 2010.

Boyens, C., Krishnan, R., Padman, R. (2004). On privacy-preserving access to distributed heterogeneous healthcare information. *In Proceedings of the 37th Annual Hawaii International Conference on Digital Object*. 2004.

Burkom, H. (2007). Alerting algorithms for biosurveillance. *In Disease surveillance: A public health informatics approach*, J.S. Lombardo and D.L. Buckeridge, Editors. 2007; Wiley: Hoboken, NJ. p. 143-192.

Cahill, C., Canales, C., Le Van Gong, H., Madsen, P., Maler, E., & Whitehead, G. (2008). Liberty Alliance Web Services Framework: A Technical Overview. Liberty Alliance Project, New Jersey. http://www.projectliberty.org/liberty/resource_center/papers, last retrieved November 2010.

CEN 13606. (2006). HealthInformatics – Electronic Health Record communication, Part1: Reference Model. *European Standard*.

Chen, Y., Wang, J. (2004). A Review of Data Integration. *Computer Science*. 2004, 31(5): 48－51.

CRISP-DM. (2010). Cross Industry Standard Process for Data Mining. http://www.crisp-dm.org/Process/index.htm, last retrieved November 2010.

Domingo-Ferrer, J. (2002). A Provably Secure Additive and Multiplicative Privacy Homomorphism. In A.H. Chan and V. Gligor (Eds.): *ISC 2002*, LNCS 2433, pp. 471–483, 2002.

Domingo-Ferrer, J. (2007). A Survey of Inference Control Methods for Privacy-Preserving Data Mining. *In Privacy-Preserving Data Mining: Models and Algorithms*. Advances in Database Systems, 2008, Volume 34, 53-80, DOI: 10.1007/978-0-387-70992-5_3.

Domingo-Ferrer, J. and Torra, V. (2001). A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 111–134, Amsterdam, 2001

Duncan, G. T. and Mukherjee, S. (2000). Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise. *Journal of the American Statistical Association*, 95:720–729, 2000.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R. and Roehrig, S. F. (2001). Disclosure limitation methods and information loss for tabular data. In P. Doyle, J. I. Lane, J. J. Theeuwes, and L. V. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 135–166, Amsterdam, 2001.

Durham, A. E., Xu, Y., Kantarcioglu, M., Malin, B. (2010). Private Medical Record Linkage with Approximate Matching, *American Medical Informatics Association Annual Symposium* 2010.

El Emam, K., Hu, J., Samet, S., Gaudette, L., Peyton, L., Earle, C., Layaraman, G., Wong, T. (under review). Secure Computation Across Health Data Registries, submitted to *Journal of the American Medical Informatics Association.* 2010.

El Emam, K., Hu, J., Mercer, J., Peyton, L., Kantarcioglu, M., Malin, B., Buckeridge, D., and Samet, S. (under second review). A Protocol for Protecting the Identity of Providers When Disclosing Data for Disease Surveillance, submitted to *Journal of the American Medical Informatics Association* (JAMIA), 2010.

El Emam, K. (2010). Risk-based health data de-identification. In IEEE Security and Privacy, 8(3):64-67, 2010.

El Emam, K. & Fineberg, A. (2009). An Overview of Techniques for de-identifying Personal Health Information. http://www.ehealthinformation.ca/documents/DeidTechniques.pdf, last retrieved November 2010.

El Emam, K. & Dankar, F. (2008). Protecting privacy using k-anonymity. *In the Journal of the American Medical Informatics Association*, September/October, 15:627-637.

El Emam, K., Jabbouri, S., Sams, S., Drouet, Y. & Power, M. (2006). Evaluating common de-identification heuristics for personal health information. *Journal of Medical Internet Research*, 8(4): e28.

ElGamal T. (1985). A prublic key cryptosystem and a signature scheme based on discrete logarithms. *In Advances in Cryptology (CRYPTO '84)*, vol. 196 of Lecture Notes in Computer Science, pp. 10–18, Springer, New York, NY, USA, 1985.

Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. (2002). Privacy preserving mining of association rules. *In Proceedings of The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 217–228). Edmonton, Alberta, Canada.

Eycken, E. V., Haustermans, K., Buntinx, F., Ceuppens, A., Weyler, J., Wauters, E., Oyen, H. V., Schaever, M. D., den Berge, D. V., Haelterman, M.. (2000). Evaluation of the encryption procedure and record linkage in the belgian national cancer registry, *Archives of public health 58* (2000), 281-294.

Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining To Knowledge Discovery: An Overview. *In Advances In Knowledge Discovery And Data Mining* , eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 1-34.

Fellegi I, Sunter A. A theory for record linkage. (1969). *J Amer Stat Assoc*. 1969; 64: 1183–210.

Fitzgerald, P. (Ed.). (2003). Interim research report. *HealthConnect. Volume 2.* Australia. http://www.health.gov.au/internet/hconnect/publishing.nsf/Content/43598FE37A3E7270CA257128007B7EB7/$File/v2.pdf, last retrieved November 2010.

Fouque, P. A., Poupard, G., and Stern, J. (2000). Sharing Decryption in the Context of Voting or Lotteries. *In Financial Crypto '00*, LNCS. Springer-Verlag.

Franklin, M. K. & Reiter, M. K. (1997). Fair Exchange with a semi-trusted Third Party. In *Proceedings of the 4th ACM Conference on Computer and Communications Security*. T. Matsumoto, Ed. Zurich, Switzerland, 1-6.

Friedrich, A.(2010). IBM Entity Analytic Solutions, IBM DB2 Anonymous Resolution: Knowledge discovery without knowledge disclosure, *IBM DB2 Anonymous Resolution Whitepaper*, ftp://ftp.software.ibm.com/software/data/pubs/papers/db2anonymousres.pdf, last retrieved May 2010.

Galindo, D and Verheul, E. R. (2010). Pseudonymized Data Sharing. *In Privacy and Anonymity in Information Management Systems*. Advanced Information and Knowledge Processing, 2010, Volume 0, Part 3, 157-179, DOI: 10.1007/978-1-84996-238-4_8

Garfinkel, R., Gopal, R. and Rice, D. (2006). New approaches to disclosure limitation while answering queries to a database: protecting numerical confidential data against insider threat

based on data and algorithms. *In Proceedings of the 39th Annual Hawaii International Conference on HICSS '06*.

Gilburd, B., Schuster, A., Wolff, R. (2004). k-TTP: a new privacy model for large-scale distributed environments. *In Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 563–568. ACM Press, New York (2004)

Grannis S, Overhage J, McDonald C. (2002). Analysis of identifier performance using a deterministic linkage algorithm. *In Proc AMIA Symp*. 2002: 305.

Goldreich, O. (2000). Modern Cryptography, Probabilistic Proofs, and Pseudorandomness, *Algorithms and Combinatorics*. ISBN 3-540-64766-x, Springer-Verlag, Vol 17, 1998.

Goldwasser S. and Micali S. (1884). Probabilistic encryption. *Journal of Computer and System Sciences*, vol. 28, no. 2, pp. 270–299, 1984.

Gopal, R., Garfinkel, R. and Goes, P. (2002). Confidentiality via camouflage: the cvc approach to disclosure limitation when answering queries to databases. *Operations Research*. 50:501–516, 2002.

Gopal, R., Goes, P., and Garfinkel, R. (1998). Interval protection of confidential information in a database. *INFORMS Journal on Computing*, 10:309–322, 1998.

Schadow,G., Grannis, J. S., McDonald J. (2002). Discussion Paper: Privacy-Preserving Distributed Queries for a Clinical Case Research Network. *IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining*. 2002.

Han, J. & Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd ed.). Morgan Kaufmann Publishers.

Hanson, J. J. (2009). Mashups: Strategies for the Modern Enterprise. Addison-Wesley, Pearson Education.

Hearst, M. A. (1999). Untangling Text Data Mining. *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

Javier Herranz, J., Nin, J. and Torra, V. (20101). Distributed Privacy-Preserving Methods for Statistical Disclosure Control. *In Data Privacy Management and Autonomous Spontaneous Security. Lecture Notes in Computer Science*, 2010, Volume 5939/2010, 33-47, DOI: 10.1007/978-3-642-11207-2_4.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, vol. 28, no. 1, pp. 75-105.

HIPAA. (1996). Health Insurance Portability and Accountability Act. United States Congress, United States. http://aspe.hhs.gov/admnsimp/pl104191.htm, last retrieved November 2010.

HL7. (2010). *Health Level Seven*. From http://www.hl7.org, last retrieved November 2010.

HL7 CDA. (2010). *Clinical Document Architecture* (CDA). http://xml.coverpages.org/healthcare.html, last retrieved November 2010.

Hu, J. & Peyton, L. (2010). A Framework for Privacy Assurance and Ubiquitous Knowledge Discovery in Health 2.0 Data Mashups, *Ubiquitous Health and Medical Informatics*, S.

Mohammed, J. Fiaidhi (Eds.), ISBN13: 9781615207770. Pages 64-83. IGI Global, Hershey, PA, USA.

Hu, J. & Peyton, L. (2009). Integrating Identity Management with Federated Healthcare Data Models. In *Proceedings of the 4th International MCeTech Conference on eTechnologies* (pp. 100-112). Ottawa, Canada. LNBIP 26, Springer.

Hu, J., Peyton, L., Turner, C., Bishay, H. (2008). A model of trusted data collection for knowledge discovery in B2B networks. *In Proceedings of the 2008 International MCETECH Conference on e-Technologies* (pp. 60-69). Montreal, Canada.

HREB. the Health Research Ethics Board, http://www.umanitoba.ca/faculties/medicine/research/ethics/, last retrieved November 2010.

Hylton, J. A. (1996). Identifying and Merging Related Bibliographic Records, Master's thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1996.

Inan, A., Saygın, Y., Sava, E., Hintoglu, A. A., Levi, A. (2006). Privacy Preserving Clustering on Horizontally Partitioned Data. *Proceedings of the 22nd International Conference on Data Engineering Workshops* (ICDEW'06).

Inan, A., Kantarcioglu, M., Scannapieco, M., Bertino, E. (2008). A hybrid approach to private record linkage. *In Proceedings of the 24th Int'l Conference on Data Engineering* – ICDE'08, 2008.

Inmon, W. H. (2005). Building the Data Warehouse, fourth Edition. Wiley Publishing.

ISDS. (2010). http://www.isdsdistribute.org/, last retrieved November 2010.

Jonas, J. (2006). Threat and Fraud Intelligence, Las Vegas Style. *Security & Privacy Magazine*, IEEE. 4, 28-34 ,2006.

Jun, J.B., Jacobson, S.H., & Swisher, J.R. (1999). Application of discrete-event simulation in health care clinics: A Survey. *Journal of the Operational Research Society*, 50(2), 109-123.

Kantarcioglu, M., Jiang, W., and Malin, B. (2008). A Privacy-Preserving Framework for Integrating Person-Specific Databases, *Privacy in Statistical Databases*, 2008, LNCS 5262, pp. 298–314.

Kantarcioglu, M., Inan, A., Jiang, W., Malin, B. (2009). Formal anonymity models for efficient privacy-preserving joins, *Data Knowl. Eng*. (2009), doi :10.1016/j.datak.2009.06.011

Karakasidis, A., Verykios, V. S. (2009). Privacy preserving record linkage using phonetic codes. *In Proceedings of the 4th Balkan Conference in Informatics*, Thessaloniki, Greece, 2009, pp. 101-106.

Kargupta, H., Datta, S., Wang, Q., and Sivakumar, K.. (2004). Random Data Perturbation Techniques and Privacy Preserving Data Mining. *Knowledge and Information Systems Journal*, volume 7, number 4, pages 387—414.

Karr, A., Lin, X., Sanil, A., Reiter, J. (2009). Privacy-preserving analysis of vertically partitioned data using secure matrix products. *Journal of Official Statistics*. 25(1), 125–138.

Kienzle, D. M., Elder, M. C., Tyree, D., Edwards-Hewitt, J. (2010). Security Patterns Repository. http://www.scrypt.net/~celer/securitypatterns/repository.pdf, last retrieved November 2010.

Kimball, R., Reeves, L., Ross, M., & Thornthewaite, W. (1998). The Data Warehouse Lifecycle Toolkit. ISBN 0-471-25547-5. John Wiley & Sons, Inc.

Koch, M., & Möslein, K.M. (2005). Identity Management for Ecommerce and Collaborative Applications. *International Journal of Electronic Commerce*, 9(3), 11–29.

Landau, S. (Ed.). (2003). Liberty ID-WSF and Privacy Overview, version 1.0, Liberty Alliance Project.
http://www.projectliberty.org/resource_center/specifications/liberty_alliance_id_wsf_2_0_specifications, last retrieved November 2010.

Le Strat, Y. (2005). Overview of temporal Surveillance. *In Spatial and syndromic surveillance for public health*, A.B. Lawson and K. Kleinman, Editors. 2005; Wiley: Chichester. p. 13-18.

Ledbetter, C.S., & Morgan, M.W. (2001). Toward Best Practice: Leveraging the Electronic Patient Record as a Clinical Data Warehouse. *Journal of Healthcare Information Management*, 15(2). http://www.himss.org/content/files/jhim/15-2/him15205.pdf, last retrieved November 2010.

Li, J. & Shaw, M. (2004). Protection of Health Information in Data Mining. *International Journal of Healthcare Technology and Management*, 6(2), 210-222.

Live. (2006). *Introduction to Windows Live ID, Windows Live Development Center*. From http://msdn.microsoft.com/en-us/library/bb288408.aspx, last retrieved November 2010.

March. S. T., and Smith, G. F. (1995). Design and Natural Science Research on Information Technology. *Decision Support Systems*, vol. 15, no. 4, pp. 251-266, 1995.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. 1. University of California Press. pp. 281–297.

MCHP. Manitoba Center for Health Policy,
http://www.umanitoba.ca/faculties/medicine/units/mchp/resources/access/proposals.html last retrieved November 2010.

Nabar S., Marthi B., Kenthapadi K., Mishra N., Motwani R. (2006). Towards Robustness in Query Auditing. *VLDB Conference*, 2006.

Naor, M., Yung, M. (1989). Universal one-way hash functions and their cryptographic applications. In *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, Seattle, Washington, United States , pp 33 – 43, 1989, ISBN:0-89791-307-8.

Neame, R.L., & Olson, M. (1996). Measures implemented to protect personal privacy for an on-line national patient index: a case study. *Topics in Health Information Management*, *17*(2),18-25

OHREB (2011) . Ottawa Hospital Research Institute Policies and Procedures.
http://www.ohri.ca/ohreb/policies.htm, last retrieved May 2011.

Ozsu, M. T., Valduriez, P. (1999). Principles of Distributed Database Systems, 2nd edition. Prentice-Hall, Upper Saddle River, NJ ,1999.

OpenEHR. (2010). *The openEHR Foundation*. http://www.openehr.org/shared-resources/getting_started/openehr_primer.html, last retrieved November 2010.

OpenSSO. (2010). *Oracle OpenSSO*. http://www.oracle.com/technetwork/testcontent/opensso-091890.html, last retrieved November 2010.

O'Reilly T. (2005). What is Web 2.0: design patterns and business models for the next generation of Software. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, last retrieved November 2010.

Paillier P. (1999). Public-key cryptosystems based on composite degree residuosity classes. *EUROCRYPT'99*.

Chaoyi Pang, C., Gu, L., Hansen, D. and Maeder, A. (2009). Privacy-Preserving Fuzzy Matching Using a Public Reference Table. *In Intelligent Patient Management*, SCI 189, pp. 71–89. Springer-Verlag Berlin Heidelberg.

Patrick, J., Wang, Y., & Budd, P. (2007). An automated system for conversion of clinical notes into snomed clinical terminology. In *Proceedings of the fifth Australasian symposium on ACSW frontiers* (pp. 219–226). Darlinghurst, Australia: Australian Computer Society, Inc.

Peyton, L., Doshi, C., & Seguin, P. (2007). An audit trail service to enhance privacy compliance in federated identity management. *In Proceedings of the 2007 CASCON conference* (pp. 175-185). Toronto, ON.

Peyton, L. & Hu, J. (2010). Federated Identity Management to Link and Protect Healthcare Data, *International Journal of Electronic Business* (IJEB). 8(3). Inderscience Publishers.

Peyton, L., & Hu, J. (2007). Knowledge discovery in a circle of trust. *In* Proceedings of Data Mining & Information Engineering 2007. Ashurst, UK: WIT Press.

Peyton, L., Hu, J. Doshi, C., & Seguin, P. (2007, July), Addressing privacy in a federated identity management network for e-health, *In Proceedings* of *Eighth World Congress on the Management of eBusiness* (pp. 12) Toronto, ON.

PHIPA. (2004), Personal Health Information Protection Act. Government of Ontario, Canada. From http://www.e-laws.gov.on.ca/html/statutes/english/elaws_statutes_04p03_e.htm, last retrieved November 2010.

Pinkas B. (2002). Cryptographic Techniques for Privacy-Preserving Data Mining. *ACM SIGKDD Explorations*, 4(2), 2002.

PIPEDA. (2000). The Personal Information Protection and Electronic Documents Act. Department of Justice, Canada. http://laws.justice.gc.ca/en/P-8.6/text.html, last retrieved November 2010.

Pseudonymization. (2008). Health Informatics – Pseudonymization. I. O. for Standardization. ISO/TS 25237:2008. http://www.iso.org/iso/catalogue_detail.htm?csnumber=42807 last retrieved November 2010

Public Health Surveillance. (2010). http://dsol-smed.phac-aspc.gc.ca/dsol-smed/ndis/glossa-eng.php, last retrieved November 2010.

Ravikumar, P., Cohen, W.W., Fienberg, S.E. (2004). A secure protocol for computing string distance metrics. *In PSDM held at ICDM*, pp. 40–46 (2004)

Recordon, D. & Reed, D. (2006), OpenID 2.0: a platform for user-centric identity management. In *Proceedings of the Second ACM Workshop on Digital Identity Management* (pp. 11-16). Alexandria, Virginia, USA. DIM '06. ACM, New York, NY, 11-16.

Rivest, R., Adleman, L., and Dertouzos, M. (1978). On data banks and privacy homomorphisms. *In Foundations of Secure Computation*, pp. 169–177, Academic Press, 1978.

Rivest R., Shamir A., and Adleman L. (1978). A method for obtaining digital signatures and public key cryptosystems. *Communications of the ACM*, vol. 21, no. 2, pp. 120–126, 1978.

Rouault, J & De Clercq, J. (2004). Identity Management Architectures. July 2004, HP Dev Resource Central.

Samet, S. and Miri, A. (2009). Privacy-Preserving Bayesian Network for Horizontally Partitioned Data. *In Proceeding of the 2009 IEEE International Conference on Information Privacy, Security, Risk and Trust (PASSAT2009)*. Vancouver, Canada, August 2009, pp. 9–16.

Sandhu, R. S., Coyne, E. J., Feinstein, H. J., & Youman, C. E. (1996). Role-based access control models. *IEEE Computer*. Vol.29, No.2, Feb., 1996. p38–47.

Sarma, A., Dong, X., & Halevy, A. (2008). Bootstrapping pay-as-you-go data integration systems. *In Proceedings of ACM SIGMOD International Conference on Management of Data*.

Scannapieco, M., Figotin, I., Bertino, E., Elmagarmid, A.K. (2007). Privacy preserving schema and data matching. In: Chan, C.Y., Ooi, B.C., Zhou, A. (eds.) *SIGMOD Conference*, pp. 653–664. ACM, New York (2007)

Schnell, R., Bachteler, T., Reiher, J. (2009). Privacy-preserving record linkage using bloom filters. *BMC Medical Informatics and Decision Making,* 9(1), 41.

Alexander Schwinn, Joachim Schelp. (2005). Design patterns for data integration. *Journal of Enterprise Information Management*, Vol. 18 Iss: 4, pp.471 – 482.

Shamir, A. (1979). How to share a secret. *Communications of the ACM*, 1979; 22(11):612-613.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing*, 5(4), 13-22.

Sokolova, M., El Emam, K., & et al (2009). Personal health information leak prevention in heterogeneous texts. *Biomedical Information Extraction International Workshop, held jointly with the 7th International Conference on Recent Advances in Natural Language Processing*.

Stolba, N., Banek, M., Tjoa, A.M. (2006). the Security Issue of Federated Data Warehouses in the Area of Evidence-Based Medicine. *In the First International Conference on Availability, Reliability and Security*, pp. 11--22. IEEE Press, Washington, USA, 2006.

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):571–588, 2002.

Sweeney. L. (2002). k-anonimity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 10(5):557–570, 2002.

Swire, P. (2010). Research Report: Application of IBM Anonymous Resolution to the Health Care Sector. http://www.ehcca.com/presentations/cclf3/swire_s5_t4.pdf, last retrieved November 2010.

Tan, LT. (1995). National patient master index in Singapore. *International Journal of Bio-Medical Computing, 40(2),* 89-93.

Tatemura J., Sawires, A., Po, O., Chen, S., Candun, K., Agrawal, D., Goveas, M. (2007). Mashup feeds: continuous queries over Web services. *In Proceedings of ACM SIGMOD International Conference on the Management of Data* (pp. 1128-1130). Beijing, China.

Thomas, R. K. (1997). Team-based Access Control (TMAC): A Primitive for Applying Role-based Access Controls in Collaborative Environments. *In Proceedings of the second ACM workshop on Role-based access control*, November 1997.

Thomsen, E. (2002). OLAP Solutions: Building Multidimensional Information Systems. New York: John Wiley & Sons, Inc.

Tourzan, J., Koga, Y. (Ed.). (2006). Liberty ID-WSF Web services framework overview, Version 2.0. http://www.projectliberty.org/liberty/resource_center/specifications/liberty_alliance_id_wsf_2_0_specifications_including_errata_v1_0_updates, last retrieved November 2010.

Vaidya, J. & Clifton, C. (2004). Privacy-preserving data mining: why, how, and what for?. *IEEE Security & Privacy*. New York, NY.

Walker, A. (1999). South Australia: best practice guidelines for patient master index maintenance. *Health Information Management*, *29*(1),43-45.

Wason, T. (Ed.). (2005). Liberty ID-FF architecture overview, version 1.2. Retrieved May 2010, from http://www.projectliberty.org/liberty/resource_center/specifications/liberty_alliance_id_ff_1_2_specifications, last retrieved November 2010.

Willenborg L., de Waal, T. (2001). Elements of statistical disclosure control: Springer-Verlag 2001.

W3C Working Group, Web Services Architecture, Note 11 February 2004, http://www.w3.org/TR/ws-arch, last retrieved November 2010.

Yakout, M., Atallah, M.J., Elmagarmid, A.K. (2009). Efficient private record linkage. In: ICDE, pp. 1283–1286. IEEE, Los Alamitos.

Yao, A. C. (1982). Protocols for secure computation. *In Proceedings of the 23rd IEEE Symposium on Foundations of Computer Science*, pages 160–164. IEEE, 1982.

Yoshioka, N., Washizaki, H., Maruyama, K. (2008). A survey on security patterns, Progress in Informatics, No. 5 pp. 35-47.