In proceedings of the 10th International Conference on Mathematics and Computing ICMC 2024

# Secure and Privacy-Compliant Data Sharing: An Essential Framework for Healthcare Organizations

Redwan Walid[1], Karuna Pande Joshi[1], and Lavanya Elluri[2]

[1] Department of Information Systems
University of Maryland, Baltimore County, Maryland, USA
{rwalid1,kjoshi1}@umbc.edu
[2] Subhani Department of Computer Information Systems
Texas A&M University-Central Texas, Texas, USA
elluri@tamuct.edu

**Abstract.** Data integration from multiple sources can improve decision-making and predict epidemiological trends. While there are many benefits to data integration, there are also privacy concerns, especially in healthcare. The Health Insurance Portability and Accountability Act (HIPAA) is one of the essential regulations in healthcare, and it sets strict standards for the privacy and security of patient data. Often, data integration can be complex because different rules apply to different companies. Many existing data integration technologies are domain-specific and theoretical, while others rigorously adhere to unified data integration. Moreover, the integration systems do not have semantic access control, which causes privacy breaches. We propose a framework using a knowledge graph for sharing and integrating data across healthcare providers by protecting data privacy. We use an ontology to provide Attribute-Based Access Control (ABAC) for preventing excess or unwanted access based on the user attributes or central organization rules. The data is shared by removing sensitive attributes and anonymizing the rest using k-anonymity to strike a balance between data utility and secret information. A metadata layer describes the schema mapping to integrate data from multiple sources. Our framework is a promising approach to data integration in healthcare, and it addresses some of the critical challenges of data integration in this domain.

**Keywords:** Privacy · Security · Electronic Health Record (EHR) · Knowledge Graph (Ontology) · Attribute-Based Access Control (ABAC) · k-anonymity · metadata

## 1 Introduction

The value of data in scientific experiments within healthcare is immeasurable, especially as the amount of data being generated continues to multiply. Different entities often create this data independently, and integrating it is crucial to

transforming it into useful information. The integration allows organizations to combine data from various sources, providing consumers with real-time insight into business outcomes, increasing productivity, improving decision-making, and enabling future predictions. While data reside in separate locations, integration is the first step toward converting data into valuable and relevant information. According to [1], the data integration system aims to provide a uniform query interface to many data sources. It can help the user from having to locate each source, interact with each source in isolation, and then combine the results.

Although data integration has numerous benefits, it also presents several challenges related to data privacy. Data privacy refers to an individual's right to control data exchange with third parties within a computer network. Data integration-related privacy concerns can be significant. In the healthcare industry, for instance, the Health Insurance Portability and Accountability Act (HIPAA) is a crucial privacy regulation that safeguards patient data in the United States [2]. Other domains have similar privacy regulations, such as the General Data Protection Regulation (GDPR), which restricts what companies can do with customer information [3]. Data providers establish security protocols that define specific criteria for acquiring, analyzing, and sharing personal and confidential data during integration. If integration systems fail to implement these security protocols properly, data leakage risks and other attacks may occur, resulting in substantial legal and financial consequences. It is, therefore, essential to implement security measures that determine who can access data for specific purposes and consider the implications of exposing such data to users.

Integrating data from various databases can be complicated due to different policies that apply across multiple organizations. The process typically involves identifying data sources, creating a mediated schema, mapping data, formulating and executing queries, and displaying the results. Nevertheless, integration systems often fail to consider users' access levels based on their attributes or roles. For example, a senior doctor may have read-and-write access to all fields in a patient's Electronic Health Record (EHR), while a junior doctor may only have read access to specific fields. While data integration technologies are available, they are often domain-specific and theoretical, and some only strictly adhere to unified data integration. Currently, no standard model in the literature considers data privacy during integration. Although many scholars have been working on data integration, only a few have addressed the challenge of data privacy.

### 1.1 Motivation

There have been various applications in the physical world that involve data integration while following strict privacy constraints. One such application is described below.

Analysis of disease occurrence, frequency, and risk factors is essential to identifying and managing them. These analyses have an enormous effect on policy decisions. An apparent precondition for such analyses is the availability of the necessary data. The data must then be gathered and integrated by various healthcare facilities while sanitizing information vulnerable to privacy.

Privacy concerns are a significant barrier to standardizing such activities. Infringement of privacy can inflict severe physical and mental distress on entities. The privacy-conserving integration and collaboration of experimental data in the health sciences have become critical to encouraging scientific exploration.

Considering the above application scenario, this work proposes a framework for sharing and integrating data across healthcare providers by protecting data privacy.

### 1.2 Our Approach

The proposed framework ensures safe sharing and data integration between various sources using different components, as listed below. An organization can exchange user data with other organizations, guaranteeing no violation of privacy.

1. The system uses HIPAA-compliant knowledge graph that uses W3C Web Ontology Language (OWL) [4] to provide Attribute-Based Access Control (ABAC) [5]. The knowledge graph helps control data leakage and provides limited user access based on user attributes and queries.
2. The data sources remove the sensitive attributes and then anonymize the data using k-anonymity [6]. K-anonymity is a technique developed to mitigate the possibility of anonymized data being re-identified when connecting to other datasets. The main objective is to anonymize the data shared by the data sources to protect privacy.
3. The system built a metadata layer on top of the data sources. The layer contains information about canonical names and names used in the local databases. The layer is used for data integration from the databases for semantic mapping. For any query, the data is derived by running sub-queries from the origins and compiling them into a unified, coherent data set.

The remainder of the paper is structured as follows – we discuss related work in Section 2, System Overview in Section 3, Dataset Description in Section 4, Implementation in Section 5, and conclusion in Section 6.

## 2 Related Work

There have been several works done to integrate healthcare data. In [7], Bahga et al. developed a framework to integrate data from distributed and heterogeneous sources into a standard terminology in the cloud environment. Their approach does not consider privacy issues. In [8], Clifton et al. proposed a privacy framework for data integration. They identified a few potential research directions and challenges that must be considered to perform privacy-preserving data integration. In [9], the authors proposed a framework in a development state, and they also identified a few research directions. Lu et al. in [10] proposed privacy solutions for smart healthcare systems. They examined identification, access control,

and detection research practices and concentrated on EHR's critical roles and features. In [11], Yau et al. presented a repository protecting privacy to integrate data from multiple data storage providers. The repository receives only the necessary amount of information from data sources depending on user requests for integration and rejects other claims. The limitation of their work is that they solely focused on matching operations.

Database researchers have also been scaling up development efforts to make privacy a core concern; for example, the Hippocratic project databases [12] and Platform for Privacy Preferences (P3P) [13]. When data is shared across various entities and transformed and integrated with other data sources, none of these initiatives answer privacy issues. In [14], the authors have suggested using a trusted third party for privacy-preserving data integration. However, they have not specified how trustworthy is the third party or if there is an evaluation metric. Organizations typically don't want to share data without such considerations. A variety of research efforts have been made on the development of data mining algorithms that protect privacy. One method adopts a distributed framework [15–17]; the other applies random noise to the data while retaining the distribution's underlying properties [18, 19]. These strategies presume data were integrated before their operations. More often, sources will not be amenable to give their data provided; data protection is maintained [8].

Several other works have been done on solving data integration challenges, and most are domain-specific [20–23]. Not many studies discussed policy issues that could impact data integration across various organizations. In [20], David addressed the data integration challenges in the omics platform, which is very challenging because the data is highly diversified, particularly in a drug discovery business that intrinsically pulls on many divergent data types. In [24], Alshawi et al. proposed a framework for assessing the quality and integration of patient data in the healthcare industry for Customer Relationship Management applications. In [25], Rigby et al. proposed a prototype knowledge broker that gathers and incorporates patient data from independent healthcare organizations using a software service model, providing a few solutions to enterprise-based file systems. Several data privacy-related works include [6, 26, 27]. The most popular methods are k-anonymity [6] and differential privacy [27]. Most of them consider privacy issues while sharing the data. Still, none considers a standard model or framework that concerns privacy issues with the data while integrating data from disparate sources, which might be very significant in a few domains like the healthcare industry, human resource recruiting companies, and financial institutions.

## 3 System Overview

Our proposed system has multiple layers depicted in the high-level system diagram in Figure 1. Assume three hospitals have planned to share their local databases with a research organization. The data sources are identified as one, two, and three. An example of how the framework will work is: the user queries

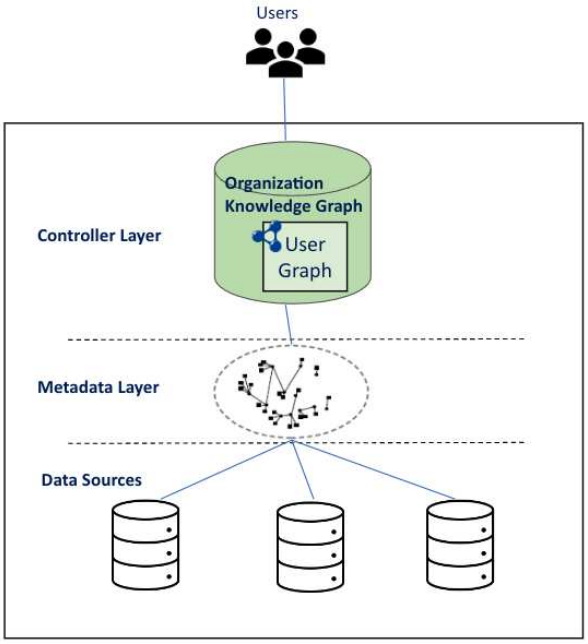


**Fig. 1.** High-level System Diagram

the framework, the user gets authenticated, and attributes get checked with a knowledge graph, the data sources remove the identifier and confidential attributes, the data sources anonymize the remaining data, subqueries get executed on the data sources, and the results are combined and displayed.

### 3.1 System Description

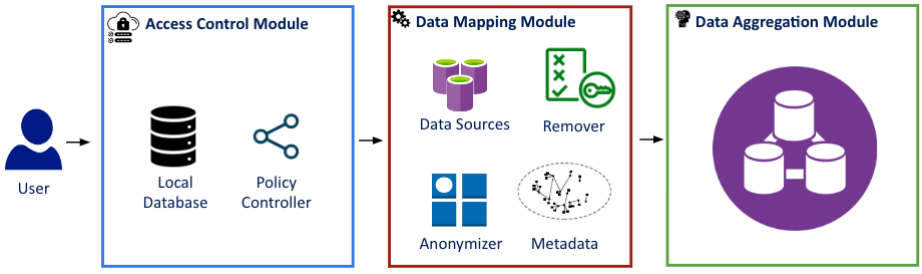


**Fig. 2.** System Architecture

The system comprises multiple components within each module, as shown in the system architecture in Figure 2. We describe each module in detail in the following few subsections.

**Access Control Module.** Attribute-Based Access Control is the primary function of the module. ABAC offers flexible, context-conscious, and cost-intelligent resource access management [28]. It allows access control strategies to require unique attributes from several separate systems to address permission and achieve successful regulatory enforcement, enabling organizations to be agile in their execution, depending on current resources. It follows Binary logic, with regulations including "IF, THEN" statements on who makes the order, the property, and the behavior.

When a user requests access, their username and password get checked with the database, and their attributes are verified. They will be denied access to the data if they do not fulfill them. Organization-defined policies are written in the Policy Controller in terms of user attributes. For example, a Senior Doctor having attributes *Certification:* CMA (Certified Medical Assistant), *Authority:* AAMA (American Association Of Medical Assistants), *Medical Institution:* GBMC (Greater Baltimore Medical Center), *Specialization:* AI (Allergy And Immunology), *Professional Designation:* Doctor will have write access to all the fields of integrated data. Likewise, a Junior Doctor will have partial access to the integrated data and may only read the data.

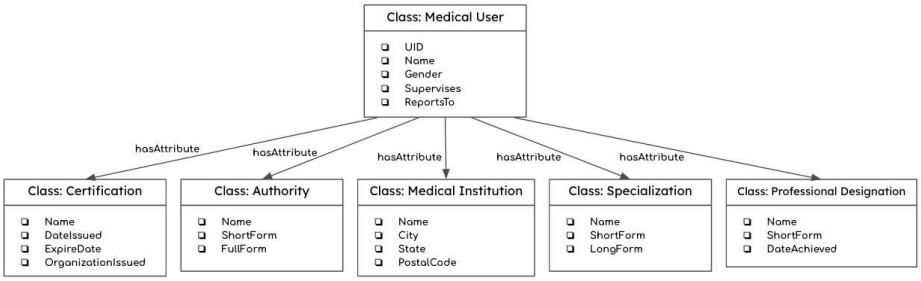


**Fig. 3.** Knowledge Graph used in the System

We use a HIPAA-compliant knowledge graph that uses W3C Web Ontology Language (OWL) [4] to control integrated data access. The knowledge graph was built by Joshi et al. [29], complying with the HIPAA standard [52], and it was transformed accordingly to capture the experimental scenario. Figure 3 shows the picture of the knowledge graph we used in our system. The knowledge graph uses the user's attribute to control access according to the policy defined by the organization. The knowledge graph may also be used to incorporate organization-specific access regulations. It can also control access to the fields of integrated data. For example, an EHR has several fields like Billing Information, Diagnosis,

Allergies, etc., and the knowledge graph can be used to control access to each field of an EHR. Likewise, in our application scenario, we use the knowledge graph to control access to the fields of the integrated data.

**Data Mapping Module.** Once a requestor passes the access control module, the data sources get identified. The data sources remove the sensitive attribute with the help of the remover. The anonymizer anonymizes the remaining data to be shared to protect privacy. Anonymization is the way of altering data until it is released for data processing. Therefore, de-identification is not achievable and would result in k identifiable information if an effort is made to de-identify by comparing the anonymized data with other data sources. We anonymize the data using k-anonymity [6] as the data utility is essential. k-anonymity is a technique developed to mitigate the possibility of anonymized data being re-identified when connecting to other datasets. k-anonymity helps to balance data utility and privacy. It is achieved by generalization and suppression.

The metadata in the module builds a layer of integration called the metadata layer above the local databases. The most common integration system uses a mediator and wrapper for integration. Mediator-Wrapper has several issues like Scalability, Flexibility, and Adaptability [56]. However, we build metadata or a semantic layer.

The metadata layer helps users independently view data using common words. It maps nuanced data into common words such as patient, age, or city to provide an organization-wide coherent, condensed view of the data. It is not a specific abstraction. It is a collection of abstractions that are used to resolve multiple issues.

We consider three databases for the experimental scenario: local DB1, local DB2, and local DB3 from data source1, data source2, and data source3. The local DBs are a collection of tables and records that we need to bring together in one database, the integrated database. The integrated database contains all the tables that make up each local DBs. The integration layer is the layer of metadata that includes information that defines the three databases. The information stored in the metadata layer can be listed as follows.

- Canonical representation: It corresponds to an object's name at the metadata layer. It is the global name of the object.
- Data type and semantic difference: It is a definition of the local data types in each database used for each column name.
- Supplementary Fields: It stores additional mapping or conversion information between canonical to local DB names.

**Data Aggregation Module.** The main task of the module involves integrating data from multiple databases into a coherent, cohesive view. In a standard data integration method, the user sends a query to the central server. The central server then collects the data needed from diverse sources. The data is derived from the origins and compiled into a unified, coherent set of data. Finally, it is fetched to the requested user.

### 3.2 Framework Flow

In our application scenario, we started with the user requesting access to the system. The user access request goes through a comprehensive check based on the user attribute check with the knowledge graph. Then the data shared by the sources were prepared by removing sensitive attributes and anonymizing using k-anonymity. After that, we built a metadata layer based on the shared databases. We now have the integration layer on top of the participating databases from the previous stage. There are two layers, the Global layer and the local layer. The global layer has metadata information and canonical representation. The local layer holds the local databases.

The user sends a query using the canonical names of fields in the metadata layer. The global query is broken down into subqueries, one for each participating local database. The subqueries are sent and executed by the applicable local databases. The outcomes are put together using a UNION process.

### 3.3 Use Cases

Data sharing and integration are widespread every day throughout organizations. Many organizations or tasks benefit from getting data integrated from different sources. We describe here two use cases that can benefit from our system.

1. Disease Outbreak: Early identification of infectious diseases is critical in avoiding life-threatening contagious diseases. Virulent disease attacks such as SARS and bird flu have rendered disease monitoring a significant issue. Outbreak identification performs well when integrating and analyzing several data points in real-time. Safeguarding identity disclosure by appropriate privacy-preserving data aggregation and sharing strategies would enable healthcare advancements.
2. Healthcare Research: Analysis of disease occurrence, frequency, and clinical signs is essential to identifying and managing them. These analyzes have an enormous effect on government policies. An apparent precondition for this research is the availability of the necessary data. The data then must be gathered and integrated by various healthcare facilities while sanitizing information that is vulnerable to privacy. Concerns over privacy are a significant barrier to simplifying such activities. A lack of privacy can cause serious harm to participants. Another issue with the violation of privacy is the chances of bias from apparently definitive statistical findings against different sub-groups. Often, privacy is handled by avoiding sharing rather than incorporating data processing restrictions into the mechanism. The privacy-preserving integration and sharing of research data in healthcare have become critical to encouraging scientific exploration.

## 4 Dataset Description

We used the MIMIC-III dataset [30] to produce synthetic data set for three independent sources. We assume each source represents a hospital. Each data source

has more than one thousand patient records. The number of attributes in each source is variable because some organizations may want to store more features to know their patient better. In contrast, others might prefer to store minimal information. We have eighteen attributes, like name, age, gender, Social Security Number (SSN), city, religion, insurance, diagnosis, drug, medical conditions, admission type, etc., from source one. Likewise, we have ten attributes from source two and twelve from source three. The data that are available from the sources are confidential and sensitive. If either of the entities shares the way-it-is info, this may lead to a breach of privacy.

## 5 Evaluation

We developed a proof of concept to evaluate our framework. Let's say a team of researchers at the National Insitute of Health (NIH) would like to know about a few statistics related to Covid19 cases. The team comprises several individuals. There is a senior research scientist, a few junior scientists, and many interns. The team would like to know the number of positive covid cases in each county in California. The team wants to see the patient's age groups, medical conditions, etc. To achieve this, the team asks for data from hospitals. The hospitals then agree to share the data by removing the sensitive or identifier attributes and anonymizing the data using k-anonymity. The team creates a knowledge graph and defines the user attribute in the knowledge graph along with their access pattern. The team also creates a metadata layer on top of the data shared by the hospitals. An intern has been asked to report California's average cases per county. Once an intern makes a request into the framework, the attributes get checked with the knowledge graph. After that, the intern puts a query that gets written into sub-queries and runs into data sources. The query result is combined from all the sources and displayed. If the senior researcher in the team performs the same operation, the final results contain more rows and columns as the attributes allow to get access to complete data. Likewise, a junior researcher gets more rows and columns than the intern.

## 6 Conclusion

This paper proposed a framework for privacy-preserving data sharing and integration. The framework used a knowledge graph for providing attribute-based access control. The access decisions were evaluated comprehensively based on the rules defined in the knowledge graph. The data shared was anonymized using k-anonymity. A semantic layer was created to represent the schema mappings and then populated with the metadata records. Later, data was integrated from different Local DBs based on the user query. The framework helps to share data and integrate securely, protecting individual privacy. The framework can be used for several use cases, like to study a disease outbreak.

There can be several expansions for this work. One of the challenging aspects would be to include many data sources and then check the system's performance. Obviously, with the addition of more sources, the complexities would add up in the metadata layer. Exploring and simplifying the challenges in building the metadata layer would be interesting. Often, machine learning is used these days for schema mapping. It would be a great area to explore. Another promising research direction would be incorporating access control into the integrated document's fields. Often, EHRs have this field-level access control. Other data anonymizing techniques could be explored as the data is very confidential, and at the same time, the utility is essential.

**Acknowledgements** This work has been supported by Center for Accelerated Real Time Analytics (CARTA), Office of Naval Research (ONR) under grants N00014-18-1-2453, N00014-19-WX-00568, and N00014-20-WX01704 and National Science Foundation (NSF) under grant 1955319.

## References

1. Alon Halevy, Anand Rajaraman, and Joann Ordille. Data integration: The teenage years. In *Proceedings of the 32nd international conference on Very large data bases*, pages 9–16, 2006.
2. Abigail English and Carol A Ford. The hipaa privacy rule and adolescents: legal questions and clinical challenges. *Perspectives on sexual and reproductive health*, 36(2):80–86, 2004.
3. Sanjay Sharma. *Data privacy and GDPR handbook*. John Wiley & Sons, 2019.
4. Deborah L McGuinness, Frank Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10(10):2004, 2004.
5. Vincent C Hu, David Ferraiolo, Rick Kuhn, Arthur R Friedman, Alan J Lang, Margaret M Cogdell, Adam Schnitzer, Kenneth Sandlin, Robert Miller, Karen Scarfone, et al. Guide to attribute based access control (abac) definition and considerations (draft). *NIST special publication*, 800(162):1–54, 2013.
6. Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.
7. Arshdeep Bahga and Vijay K Madisetti. Healthcare data integration and informatics in the cloud. *Computer*, 48(2):50–57, 2015.
8. Chris Clifton, Murat Kantarcioğlu, AnHai Doan, Gunther Schadow, Jaideep Vaidya, Ahmed Elmagarmid, and Dan Suciu. Privacy-preserving data integration and sharing. In *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 19–26, 2004.
9. Sourav S Bhowmick, Le Gruenwald, Mizuho Iwaihara, and Somchai Chatvichienchai. Private-iye: A framework for privacy preserving data integration. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 91–91. IEEE, 2006.
10. Yang Lu and Richard O Sinnott. Security and privacy solutions for smart healthcare systems. In *Innovation in Health Informatics*, pages 189–216. Elsevier, 2020.

11. Stephen S Yau and Yin Yin. A privacy preserving repository for data integration across data sharing services. *IEEE Transactions on Services Computing*, 1(3):130–140, 2008.
12. R Agrawal J Kiernan R Srikant and Yirong Xu. Hippocratic databases. In *Proc. 28th Int'l Conf. Very Large Databases (VLDB)*, 2002.
13. Lorrie Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph Reagle. The platform for privacy preferences 1.0 (p3p1. 0) specification. *W3C recommendation*, 16, 2002.
14. Susan W van den Braak, Sunil Choenni, Ronald Meijer, and Anneke Zuiderwijk. Trusted third parties for secure and privacy-preserving data integration and sharing in the public sector. In *Proceedings of the 13th Annual International Conference on Digital Government Research*, pages 135–144, 2012.
15. Yehuda Lindell and Benny Pinkas. Privacy preserving data mining. In *Advances in Cryptology—CRYPTO 2000: 20th Annual International Cryptology Conference Santa Barbara, California, USA, August 20–24, 2000 Proceedings*, pages 36–54. Springer, 2000.
16. Wenliang Du and Mikhail J Atallah. Secure multi-party computation problems and their applications: a review and open problems. In *Proceedings of the 2001 workshop on New security paradigms*, pages 13–22, 2001.
17. Murat Kantarcioglu and Chris Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE transactions on knowledge and data engineering*, 16(9):1026–1037, 2004.
18. Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 439–450, 2000.
19. Hillol Kargupta, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *Third IEEE international conference on data mining*, pages 99–106. IEEE, 2003.
20. David B Searls. Data integration: challenges for drug discovery. *Nature reviews Drug discovery*, 4(1):45–58, 2005.
21. David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. Data integration in the era of omics: current and future challenges. *BMC systems biology*, 8(2):1–10, 2014.
22. Nesime Tatbul. Streaming data integration: Challenges and opportunities. In *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, pages 155–158. IEEE, 2010.
23. Vladimir Gligorijević and Nataša Pržulj. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112):20150571, 2015.
24. Sarmad Alshawi, Farouk Missi, and Tillal Eldabi. Healthcare information management: the integration of patients' data. *Logistics Information Management*, 16(3/4):286–295, 2003.
25. David Budgen, Michael Rigby, Pearl Brereton, and Mark Turner. A data integration broker for healthcare systems. *Computer*, 40(4):34–41, 2007.
26. Elena Zheleva and Lise Getoor. Privacy in social networks: A survey. *Social network data analytics*, pages 277–306, 2011.
27. Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings 5*, pages 1–19. Springer, 2008.

28. Eric Yuan and Jin Tong. Attributed based access control (abac) for web services. In *IEEE International Conference on Web Services (ICWS'05)*. IEEE, 2005.
29. Karuna Pande Joshi, Yelena Yesha, Tim Finin, et al. An ontology for a hipaa compliant cloud service. In *4th International IBM Cloud Academy Conference ICACON 2016*, 2016.
30. Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.