frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 03 March 2023
DOI 10.3389/fpubh.2023.1125011

An anonymization-based privacy-preserving data collection protocol for digital health data

J. Andrew[1]*, R. Jennifer Eunice[2] and J. Karthikeyan[3]*

[1]Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India, [2]Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India, [3]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India

OPEN ACCESS

EDITED BY
Steven Fernandes,
Creighton University, United States

REVIEWED BY
Sidra Abbas,
COMSATS University Islamabad, Sahiwal Campus, Pakistan
Jatinderkumar R. Saini,
Symbiosis Institute of Computer Studies and Research (SICSR), India

*CORRESPONDENCE
J. Andrew
✉ andrew.j@manipal.edu
J. Karthikeyan
✉ karthikeyan.jk@vit.ac.in

SPECIALTY SECTION
This article was submitted to Digital Public Health, a section of the journal Frontiers in Public Health

RECEIVED 15 December 2022
ACCEPTED 06 February 2023
PUBLISHED 03 March 2023

CITATION
Andrew J, Eunice RJ and Karthikeyan J (2023) An anonymization-based privacy-preserving data collection protocol for digital health data. *Front. Public Health* 11:1125011. doi: 10.3389/fpubh.2023.1125011

Digital health data collection is vital for healthcare and medical research. But it contains sensitive information about patients, which makes it challenging. To collect health data without privacy breaches, it must be secured between the data owner and the collector. Existing data collection research studies have too stringent assumptions such as using a third-party anonymizer or a private channel amid the data owner and the collector. These studies are more susceptible to privacy attacks due to third-party involvement, which makes them less applicable for privacy-preserving healthcare data collection. This article proposes a novel privacy-preserving data collection protocol that anonymizes healthcare data without using a third-party anonymizer or a private channel for data transmission. A clustering-based k-anonymity model was adopted to efficiently prevent identity disclosure attacks, and the communication between the data owner and the collector is restricted to some elected representatives of each equivalent group of data owners. We also identified a privacy attack, known as "leader collusion", in which the elected representatives may collaborate to violate an individual's privacy. We propose solutions for such collisions and sensitive attribute protection. A greedy heuristic method is devised to efficiently handle the data owners who join or depart the anonymization process dynamically. Furthermore, we present the potential privacy attacks on the proposed protocol and theoretical analysis. Extensive experiments are conducted in real-world datasets, and the results suggest that our solution outperforms the state-of-the-art techniques in terms of privacy protection and computational complexity.

KEYWORDS

anonymization, data privacy, healthcare data, *k*-anonymity, privacy-preserving, data collection

1. Introduction

Healthcare industries have seen a significant transition since the advancements in communication technologies (1). E-health services (2) have become popular for their wide range of advantages such as accuracy, timeliness, easy access, and efficiency (3, 4). Electronic health records (EHRs) (5) are the major step toward the transformation of traditional healthcare services into paperless medical practice that can reduce the risk of medical errors (6–8). Digitized patients' health record benefits both patients and healthcare providers in sharing, monitoring, tracking, and analyzing the healthcare of patients (9). As EHRs follow a standard health record format, it is possible to make them available worldwide (10). EHRs reduce administrative overhead, costs, and medical errors through efficient communication of health information (11). Healthcare organizations often collect EHRs for medical and research purposes (12). EHRs generally contain information concerning individual health records, medical history, medications, physical conditions, etc. (13). Since there is a huge amount of personal information contained in EHRs, it is crucial to consider privacy issues more carefully (14–16).

Collecting personal health records without breaching the privacy of involved individuals is essential for its success (17–20). In the data collection problem, the data collector is usually an untrusted third-party service provider who collects data from a set of individual data owners (21, 22). Assume that a medical researcher requests data from a number of patients who hold the healthcare demographics. The schema of demography consists of user ID, age, sex, weight, and diagnosis that every patient provides to the data collector. The health record schema is a combination of personal identifiers (e.g., user ID), quasi-identifiers (QI) (e.g., age, sex, weight), and a sensitive attribute (e.g., diagnosis). A sample healthcare records collection table is shown in Table 1.

In the aforementioned example, although there are no direct identifiers such as name and social security number (SSN) in the EHR, privacy breaches can still arise. An untrusted data collector can ascertain the identity of the patient through the explicit identifier *userID* and sensitive attribute *diagnosis* of each individual. Although QI cannot be used to directly identify a person, by connecting them to the data in a published database, it may be possible to do so. The QI can act as an identifier in the absence of a direct identifier. Hence, identity disclosure is one of the major privacy issues in EHR. In the data collection problem, identity disclosure (23) can arise both at internal and external levels. Internal identity disclosure (24) generally happens within the organization either through the data owners or the data collectors. External identity disclosure (25) takes place when the data is transmitted between the owner and the collector.

Unsurprisingly, privacy-preserving healthcare data collection has become a recent research focus where a good number of literature exists (26–32). Cryptography or anonymization-based approaches are widely used to prevent the identity disclosure of EHR (33, 34). Symmetric key and asymmetric key cryptography, multiparty computation, and homomorphic encryption are some of the cryptographic approaches used for privacy-preserving data collection (35); although it guarantees privacy to a certain extent, significant challenges such as heavy computation and key propagation make it a difficult choice. The anonymization approach (36), in general, removes the identifiers and generalizes the QIs excluding the sensitive attribute. Traditional anonymization techniques, such as *k*-anonymity (37), *l*-diversity (38), *t*-closeness (39), clustering-based *k*-anonymity (40), (α, k)-anonymity (41), *p*-sensitive *k*-anonymity (42), and others, anonymize the personal records by grouping similar QI attributes to make them indistinguishable from other sets of records in the same table.

Most of the literature for privacy-preserving data collection has not considered distributed data owners, and it is assumed that personal data are already collected in a common place to be anonymized (43). Hence, in centralized solutions for privacy-preserving data collection, it has become essential to employ a third-party anonymizer (44). However, it is highly undesirable for a patient to share his/her original EHR with a third party. There is also a huge risk of a privacy breach when a data owner (patient) directly shares their personal information with the data collector. The existing privacy models drudged to control the disclosure by deploying an anonymization layer or private unidentified channel between the data collector and the data owner. Nonetheless, such assumptions are not practical as the layer or channel is not persistent. Cryptographic approaches also encrypt the healthcare records to prevent identity disclosure at the data collector's end; furthermore, the data are anonymized, resulting in poor data utility.

In this research, we propose a data collection protocol for EHRs that is effective and protects privacy in order to address the aforementioned problems. In the proposed protocol, multiple data owners anonymize their health records in a distributed and collaborative fashion before submitting the data to the data collector. This protocol's main goal is to forbid explicit exchanges between data owners and data collectors. The data owners submit their anonymized QIs through a set of representatives elected for their equivalent group. Representatives are data owners of the equivalent group with common quasi attributes. Every equivalent group should satisfy the clustering-based *k*-anonymity property (i.e., at least *k*-1 records share the same quasi attributes); therefore, the anonymized records with common QIs are submitted to the data collector through group representatives. This approach of the proposed protocol is efficient in tackling internal and external identity disclosure. Table 1 shows the original EHR of *n* patients, Table 2 shows the anonymized version of the original records by the proposed protocol. As shown in Table 2, there are two equivalent groups that share common QIs of size *k* = 3. Such equivalent groups, along with sensitive values (e.g., diagnosis), are collected by the data collector, which reduces the risk of identity disclosure. Furthermore, dynamic data owners

TABLE 1 Electronic health records.

User ID	Age	Sex	Weight	Diagnosis
1,2,3	30–40	F	55	Gastritis
		F	50	Flu
		F	60	Dyspepsia
4,5,6	55–65	M	65	Pneumonia
		M	75	Flu
		M	68	Cancer

TABLE 2 3-anonymized health records.

User ID	Age	Sex	Weight	Diagnosis
1	35	F	55	Gastritis
2	40	F	50	Flu
3	45	F	60	Dyspepsia
4	55	M	65	Pneumonia
5	60	M	75	Flu
6	65	M	68	Cancer

who join or leave an equivalent group are handled by a greedy heuristic method.

The major contributions of the proposed protocol are as follows:

(1) **Privacy-preserving healthcare data collection protocol:** A novel *k*-anonymity-based data collection protocol specifically for healthcare data collection is proposed.
(2) **Leader election:** A leader election algorithm is proposed to elect representatives of equivalent groups of anonymized records that share similar generalized quasi attributes.
(3) **Greedy heuristic method:** Data owners who dynamically join or leave the group is efficiently managed without affecting the data utility and privacy.
(4) **Leader collision mitigation and sensitive attribute protection:** We propose solutions for privacy breach through leader collision and methods to enhance the protection of sensitive attributes.

The remainder of this article is structured as follows. The recent state-of-the-art literature is discussed in the Section 2. In the Section 3, an adversarial model of the proposed protocol is presented, along with a data model and other definitions. In the Section 4, the proposed protocol is formally defined, along with the proposed algorithms. In the Section 5, data utility and possible privacy attacks on the proposed protocol are discussed. In the Section 6, experiments conducted are presented. Finally, the Section 7 concludes the article.

2. Literature survey

In the last decade, a huge number of research studies were conducted in privacy-preserving data publication and data collection. This section presents a detailed study of various state-of-the-art literature available in the field of preserving the privacy of personal data. In privacy-preserving data collection and publication, disclosure or reidentification of data owners has been a significant issue. The state-of-the-art literature consists of cryptographic and anonymization-based approaches for privacy preservation. The collection of personal data is accomplished through devices and sensors. The device periodically collects and transmits the data to the data collector upon request. The data transmission is generally conducted in a closed or open network. Hence, it is essential to ensure the secure transmission of data. Hussien et al. (45) used a symmetric key cryptographic technique to propose a secure and energy-efficient method to collect data in wireless sensor networks.

Most privacy-preserving schemes require a secure transmission channel or a third-party authentication system. However, they are impractical due to various challenges. In (46), Beg et al. have proposed a reversible data transform (RDT) algorithm for privacy-preserving data collection in the mobile recommendation system (MRS). The proposed RDT algorithm is used to protect sensitive attributes. To avoid the third-party role in the data collection process, the data transfer is done through elected representatives. However, the leader election process is straightforward, and leader collision is possible that can breach privacy. However, the same authors in (47) addressed the RDT prior data sharing and its parameter protection challenges by proposing a chaotic RDT for PPDP MRS. The authors also claim that the proposed approach can replace homomorphic encryption techniques and preserve the privacy of the MRS. The leader collusion problem is addressed by Sajjad et al. (48) through a random leader election mechanism that elects the leaders randomly and maintains a leader table for maintaining the records. However, this scheme is inefficient, which simply uses a random function to select the leaders, and leader collusion is still possible when the number of available groups is minimal. Data anonymization is vital in protecting big data and IoT data. Ni et al. (49) evaluated the performance of data anonymization schemes in an IoT environment for big data. The authors addressed the reidentification risks and evaluated the schemes based on privacy preserving-level and data utility metrics. Traditional anonymization schemes like *k*-anonymity, *l*-diversity, obfuscation, permutation, and differential privacy techniques (50) are evaluated through information loss, data utility, and conditional entropy. A similar study was presented by Sun et al. (51) for trajectory data publishing. Canbay et al. (52) proposed a Mondrian-based utility aware anonymization approach called u-Mondrian. This approach is aimed to address the upper-bound problem in the Mondrian anonymization approach that leads to poor data utility.

Healthcare data contain sensitive information that must be protected concurrently; it is very vital for healthcare research. Hence, it is essential for protecting the privacy of healthcare data with appropriate data utility. In (53), we proposed a clustering-based anonymization approach for privacy-preserving data collection in a healthcare IoT environment. The proposed approach utilizes a client–server model to anonymize the healthcare data before it reaches the data collector. The model is evaluated with information loss and other data utility metrics. A similar approach was proposed by Abbasi and Mohammadi (54) to protect the privacy of healthcare data in cloud-based systems. They proposed an optimal *k*-anonymity technique called the *k*-means++ method and used the normal distribution function to improve the anonymization data utility. We performed another study called an attribute-focused approach (55) to protect the privacy of healthcare data during data publishing. In this study, the healthcare attributes are categorized as numerical and sensitive attributes. A fixed-length interval approach is used to protect the numerical attributes and an improved *l*-diversity approach is used to protect the sensitive attributes. Avraam et al. (56) proposed a deterministic approach for protecting the privacy of sensitive attributes. This approach identifies the categorical and continuous attributes from the dataset and applies different mechanisms to prevent a privacy breach. The stratification technique is used for categorical and continuous attributes that are redistributed based on *k*-nearest-neighbor algorithms. The proposed approach is claimed to be efficient in preventing the data from reidentification. Kanwal et al. proposed multiple anonymization-based approaches to preserve the privacy of health records. In (57), they proposed a privacy scheme called horizontal sliced permuted permutation to protect multiple records of data owners. They considered the protection of multiple sensitive attributes by proposing 1: M MSA-(p, l)-diversity approach (58). Furthermore, the authors proposed an anonymization technique with an access control mechanism for hybrid healthcare cloud services. In all the studies, they evaluated data privacy for various privacy attacks such as identity disclosure attacks, membership disclosure, and sensitive attribute disclosures. Jayapradha and Prakash (59) presented a privacy-preserving model called f-slip that uses a frequency-slicing approach to protect sensitive attributes. Sensitive attributes are correlated to maintain the linking relationship during the anonymization process. Khan et al. (60) used phonetic encoding and generalization approaches for record linkage problems. The authors used phonetic encoding for anonymizing textual data, and for categorical and numerical attributes, the k-anonymization-based approach is utilized. Raju and Naresh (61) proposed a distributed algorithm to merge the datasets from different sources to maintain their privacy. To preserve the privacy of the sensitive attributes, they proposed a bucketization-based approach called $(l,m,d)^*$- anonymity. The proposed approach anonymizes the data and transforms the data into a sensitive attribute and quasi-attribute table.

Based on the in-depth literature study of the recently published literature, most of the privacy-preserving models are still using the k-anonymization-based approach. However, they either use a private secure channel or a third-party anonymizer for privacy-preserving data collection. This may lead to a possible privacy breach. Hence, a k-anonymity-based privacy-preserving protocol for data collection without a third-party anonymizer is on demand.

3. Preliminaries

Various terminologies used in this study are introduced in this section. The components of the proposed protocol such as the data model, adversary model, and system architecture are defined.

3.1. Data model

We assume that EHRs are generated periodically on the users' devices. Out of the different attributes of personal healthcare data, only the major attributes such as personal identifiers, QIs, and sensitive attributes are considered in this article. Personal identifiers are explicit attributes that unambiguously distinguish a particular individual (e.g., social security number, name, IP address, and phone number). Identifiers are generally removed in the process of data collection and publication to avoid identity and attribute disclosure.

QIs are common attributes that can be shared by more than one data owners (e.g., age, sex, and zip code). Although they cannot directly identify an individual, the combination of QIs with publicly available datasets may breach privacy. In general, generalization and suppression approaches are used to protect QIs. Sensitive attributes (S) are details about a person that should not be shared (e.g., diagnosis). Identification of an individual's sensitive information, along with the identity, is a serious privacy breach. Hence, sensitive information is needed and protected with top priority.

3.1.1. Definition 1: (Personal health data)

In personal health records table T, let H be a unique record in the table and H^{qi} be one of the QIs, and H^{si} be the single sensitive attribute (S) of the particular record. The health data schema is then defined as follows:

$$\left(H_1^{qi}, H_2^{qi}, H_3^{qi}, \ldots, H_m^{qi}, H^{si}\right)$$

where m is the number of QIs for the record. In this article, a single sensitive attribute problem is considered.

3.1.2. Definition 2: (Anonymization)

The term anonymization means protecting the identity. Hence, it involves a process of transforming the original health records to an equivalent less significant record. The original health record table T is mapped with an anonymization function f to generate an anonymized table T^*. Every record of t in T is mapped to a record in T^*. The anonymized QI attribute QI^* for every t in T^* is then defined as $t_i\,[QI] \prec\ t_i^*[QI]$.

3.1.3. Definition 3: (k-anonymity)

A personal health dataset T satisfies k-anonymity when a record t of T^* is imperceptible from at least k-1 other records. It is given by $k\ \leq\ N(t\,(QI))$ for every record $t\ \in\ T$, $N(t\,(QI))$ – number of records shares the same QI.

3.1.4. Definition 4: (Clustering-based k-anonymity)

A personal health dataset T satisfies the clustering-based k-anonymity (25) property if a set of clusters formed from n records where each cluster consists of k records where $k\ \leq\ n$.

3.1.5. Definition 5: (Equivalence class)

To create an equivalent class, at least k data owners' anonymized records with related quasi characteristics must be used. Let G^E represent the collection of data owners k who are grouped by the same anonymized quasi attributes QI^*. G^E is an equivalent group if and only if $G^E\ =\ \left\{d|d\,[QI] = qi\right\}$ and $k\ \leq\ G^E$, where d represents an arbitrary data owner with quasi attribute $d\,[QI\ ]$.

3.2. Adversary model

In privacy-preserving healthcare data collection context, there could be a single data collector and multiple data owners.

Personal health data are generated by data owners (Definition 1). We assume that there are n data owners in the network and can communicate with other data owners and the collector. The client devices (e.g., medical sensors) at the data owner's end perform communication. The data owners collaborate with other clients not only to protect their health data but also patients in the network.

The data collector collects anonymized health records from the patients. In our protocol, the data collector is assumed to be a single semi-honest collector in the network. A semi-honest entity in a network generally follows the protocols but sometimes breaches the protocol to acquire more information. An attempt may be made to learn more about a person by a semi-honest data collector. This leads to identity disclosure.

A group of data owners who share the same quasi attributes forms an equivalent group (Definition 5) satisfying the k-anonymity and clustering-based k-anonymity model (i.e., at least k data owners in an equivalent group). Table 2 shows the example of an anonymity model that contains two groups with the value of $k = 3$. The records in the equivalent group share similar quasi attributes. The data owners interact with the data collector through the equivalent groups. Thus, it protects the data from external identity disclosure. Since the data owners share common quasi attributes in an equivalent group, internal identity disclosure is also protected.

An adversarial model is necessary to identify possible privacy attacks in the system. In a privacy-preserving data collection model, an adversary could be a data collector and data owner. The data collector is considered to be a malicious component in the network. Therefore, giving the data collector access to the original records is not appropriate. The clustering-based k-anonymity model ensures anonymized data is submitted to the data collector. The data owner can also be an adversary. An adversarial data owner generates fake quasi attributes and gets added to a specific equivalent group. During the random election of group representatives, if the adversarial data owners are elected as the first and second leaders of the group, then the sensitive attributes are disclosed. Such an attack is called a leader collision attack (LCA).

3.3. Overview of the protocol

Initialization, leader election, and data collecting phases make up the proposed data collection process. In the initialization step, the data owners (patients) create QI attributes and provide them to the data collector (without sensitive attributes). The data collector applies the provided clustering-based k-anonymity model to anonymize the health records. This results in the original QI being equivalent to at least k-1 generalized quasi characteristics (GQI). The appropriate data owners are then given the GQI and the list of data owners. The data owners then create comparable groupings that comply with the privacy policy.

In the leader election phase, members of an equivalent group are assigned with unique numbers; then based on a random number generation function, two leaders are elected for each equivalent class. The first leader obtains each member's hidden sensitive attributes from the phase of data collecting that uses sensitive values that are not real. The GQI and list of sensitive data are then given to the data collector. Without actually possessing sensitive information, the second leader gathers counterfeit sensitive information. In order to obtain the anonymized dataset, the data collector then executes intersection operations on the first and second leader datasets. The proposed privacy-preserving data collection protocol's architecture is depicted in Figure 1.

The proposed approach additionally takes into consideration of dynamic data owners who join or depart the equivalent class during the anonymization process. Dynamic join or leave follows the privacy requirement and ensures the required number of members for each group.

4. Privacy-preserving healthcare data collection protocol

Initialization, leader election, and data collection are the three phases of the protocol. The anonymization network is organized during the initiation phase, and the QI properties of the data owners are generalized. Representatives from related groups were chosen to serve as the leader during the election process. The data collector is finally given access to the anonymized records

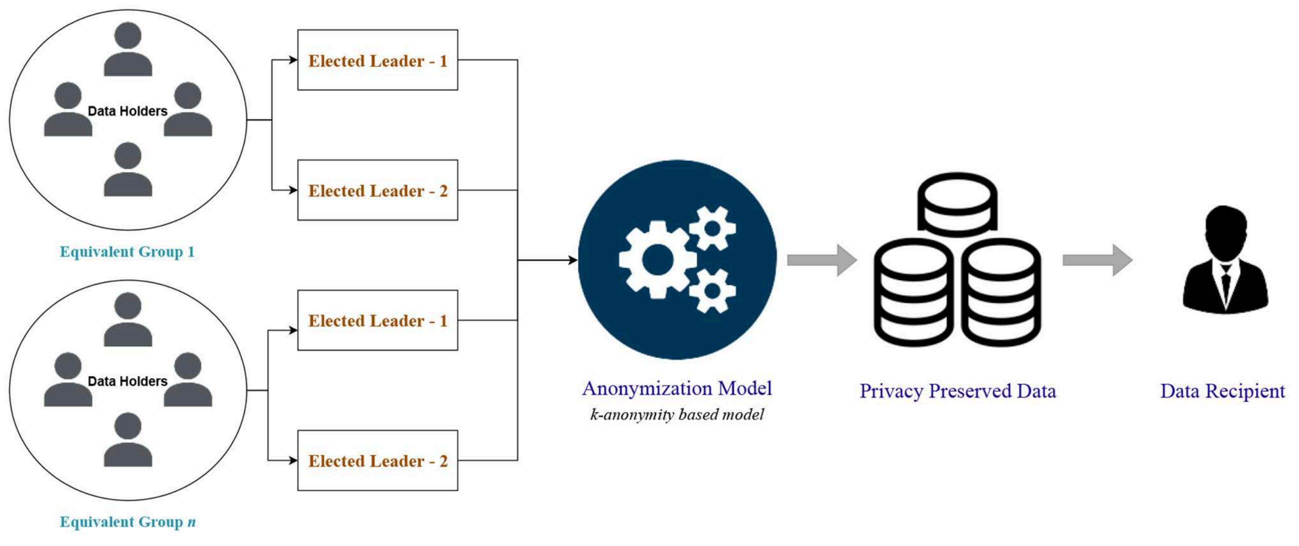

FIGURE 1
Privacy-preserving data collection protocol architecture.

with quasi characteristics and sensitive attributes during the data collecting phase. We also outline techniques for managing data owners who join or leave the network on a dynamic basis.

4.1. Initialization phase

The anonymization network is set up by the initialization phase. Data owners and data collectors are required to initialize their attributes for the network. There are two algorithms proposed for data owner initialization and data collector initialization. Data owners initially transmit their QI attributes to the data collector over the specified network. It should be highlighted that the data owners do not send their sensitive qualities. Over time, the data collector gets QI attributes from n data owners. Then the data collector anonymizes the QI attributes based on any given privacy model (37–40) to generate generalized quasi attributes (GQI). For example, Table 1 shows the original health records of n ($n = 6$) data owners that are sent to the data collector without the sensitive attribute (e.g., diagnosis). Table 2 shows the anonymized version of Table 1 with the value of $k = 3$.

The generated GQIs are distributed to the relevant data owners together with a list of data owners who have common GQIs. The list is then used by the data owners to connect with other data owners who have the same GQI. Every data owner then verifies their GQI with other data owners to form an equivalent group. Equivalent groups should satisfy the privacy policy of at least k data owner records present in every group. For example, Table 2 shows two equivalent groups that share the same GQI. The detailed steps of initialization for the data owner and data collector are shown in Algorithms 1, 2. Table 3 describes the symbols used in the algorithms.

Algorithm 1 runs at the data collector end to receive the quasi attributes from the data owners and to generate GQI based on any given anonymization techniques. It then disseminates the GQIs to the data owners. Algorithm 2 runs at the data owner's end to send the QIs to the data collector and to form equivalent groups based on the received GQI.

```
Input: QI - Data owners quasi attributes, k - privacy parameter
Output: GQI - Data collector's generalized quasi attributes
for each QI received from data owner $D_i$ do
    insert QI into QIT
end for
$G_{ID}$= Group ID
while $QIT \neq NULL$ do
    anonymize QI to GQI w.r.t k
    insert GQI into GQIT
    $G_{ID} = G_{ID} + 1$
end while
return GQIT
return D list of data owners
```

Algorithm 1. Data collector—initialization.

4.2. Leader election phase

On the data owners' side, equivalent classes are formed as per the privacy requirement k. In the leader election phase, two leaders are elected to represent the group and interact with the data collector. Algorithm 3 shows the detailed steps for leader election. First, the equivalent class members are counted. Then the *random()* function is used to generate two random numbers between 1 and the maximum number of members in the group. First, the randomly generated *userID* is considered as the first and second leader. Then we identified the energy and delay-less efficient leaders by utilizing the firefly-based algorithm proposed by Sarkar and Senthil Murugan (62). Firefly-based algorithm calculates the Euclidean distance between the elected leader and the nodes in the network then based on the distance metrics a firefly with cyclic randomization is performed to select the best leaders from among the groups. After every leader election, the leader table is updated. This algorithm ensures a single data owner is selected as the first and second leader. The elected leaders then transfer data to the data collector in the data collection phase.

4.3. Data collection phase

The major task of the data collection phase is to collect the anonymized personal health records from the data owners. During the data collection initialization stage, QI attributes of data owners are generalized by the data collector then equivalent groups are formed on the data owners' side. To avoid explicit interaction of data owners with the data collector, group leaders are elected in the leader election phase. The leaders of each group are responsible for communicating QIs and sensitive identifiers. There are two leaders elected, the first leader (L^1) is responsible to send the generalized QIs and multivalued sensitive attributes (MSA). The members equivalent group sends

```
Input: GQIT from data collector, D list of data owners
Output: $G^E$ - set of equivalent groups
for all d $\epsilon$ $D_i$ do
    generate QI
    send QI to the data collector
end for
receive GQIT, D from data collector
for all gqi $\epsilon$ GQIT do
    if $gqi_i == d(GQI)$ then
        insert GQI into $G_i^E$
        continue
    else
        break
    end if
end for
Get consent to add d in equivalent group $G_i^E$
return $G^E$
```

Algorithm 2. Data owner—initialization.

TABLE 3 Symbols.

Symbols	Description	Symbols	Description
QI	Quasi identifier	R_{GE}	Number of records in G^E
QIT	Quasi identifier table	G	Number of groups in anonymized dataset
GQI	Generalized quasi identifier	L^1	First leader
GQIT	Generalized quasi identifier table	L^2	Second leader
G_{ID}	Group ID	LT	Leader information table
D	Data owner	U^{ID}	Group member user ID
G^E	Equivalent group	CS_j	Counterfeit sensitive information of L^1
ST_R	Sensitive information of L^2	ST_j	Number sensitive information in L^1
AT	Anonymized table	S_j	Sensitive attribute in final table AT

```
Input: $G^E$- set of equivalent groups
Output: $LT$ - Leader Table with their respective group id $G_{ID}$
$R_{GE}$ - Number of records in an equivalent group $G^E$
$R_1$, $R_2$ = Values ranging from 1 to $R_{GE}$ for every $G_i^E \in G^E$
$L^1$ = rand ($R_1$, $R_2$)
$L^2$ = rand ($R_1$, $R_2$)
Calculate Euclidean distance between leader and the group members
Identify the leaders by firefly cyclic randomization (62)
if ($L^1 \neq L^2$) then
    insert $L^1$,$L^2$ into $LT$
    insert respective $G_{ID}$ into $LT$
end if
return $LT$
```

Algorithm 3. Leader election.

their anonymized records along with the multivalued sensitive attribute to the first leader. The MSA is a combination of an original sensitive attribute and n-1 counterfeit-sensitive attributes (where n is the size of the equivalent group's records). Hence, the first leader cannot discern the sensitive attributes of others in the group. Table 4 shows the example of the first leader anonymized dataset. The members of an equivalent class send their counterfeit sensitive attributes (CSA) (without the original sensitive attribute) to the second leader (L^2). Table 5 shows the example of the second leader dataset that only contains CSA along with the userID.

The data collector receives the datasets for the first and second leaders from each equivalent group during the data collecting phase. Elimination of counterfeit information from the first leader dataset is another important process for data collectors. It is hard for the data collector to identify the first and second leader datasets of each equivalent class as it performs subtraction and aggregation to eliminate the CSA from the dataset. The detailed steps of the data collection phase are given in Algorithm 4.

TABLE 4 Anonymized data collection (first leader).

User ID	Age	Sex	Weight	Diagnosis
1,2,3	30–40	F	50–60	Gastritis, heart disease, pneumonia
		F		Flu, cancer, osteoarthritis
		F		Dyspepsia, gastritis, flu
4,5,6	55–65	M	65–75	Pneumonia, cancer, arrhythmia
		M		Flu, bronchitis, pneumonia
		M		Cancer, heart disease, gastritis

TABLE 5 Anonymized data collection (second leader).

User ID	Diagnosis
1,2,3	Heart disease, pneumonia
	Cancer, osteoarthritis
	Gastritis, flu
4,5,6	Cancer, arrhythmia
	Bronchitis, pneumonia
	Heart disease, gastritis

4.4. Dynamic data collection phase

The data collection protocol is designed in a way that it can consider data owners who join or depart the network dynamically. Dynamic data owners have to be efficiently managed to avoid any privacy breach to the network. The challenges with dynamic data owners are when a dynamic data owner joins the network, he/she should be placed in the appropriate equivalent group with minimal information loss and when a dynamic data owner leaves the network it should not affect the required privacy policy and without any privacy breach. During dynamic join or leave, the entire equivalent group needs to be reorganized, which incurs huge computational costs. Hence, the greedy heuristic method is proposed to efficiently handle dynamic data owners.

```
Input: $L^1$- Dataset, $L^2$ - Dataset, $GQIT$
Output: $AT$ - Anonymized Table
$g$ = number of groups
$U^{ID}$ = user id of group $g_i$
$CS_j$ = counter sensitive attribute of $L_i^1$ at
   column $j$
$ST_R$ =sensitive information of $L_i^2$ of a
   particular $U^{ID}$
$ST_j$ = number of QIs in $L^2$
$S_j$ = sensitive attribute after removing
   counterfeit information
for $i$ = 1 to $g$ do
        for $j$ = 1 to $ST_j$ do
                if $CS_j$ == $ST_j$ then
                        $S_j = CS_j - ST_j$
                        insert $S_j$ to $AT$
                end if
        end for
end for
return $AT$
```

Algorithm 4. Data collection.

4.4.1. Dynamic join

When dynamic data owners try to join the network, they transmit the data collector their QI attributes. The data collector considers the QI attribute as a dynamic join request and finds appropriate GQI from the existing GQIT to minimize the information loss. The new data owner is then added to the particular equivalent group who the share same GQI. The representatives (the first leader and the second leader) and group members are then notified about the new member in the group along with the modified GQI. Thereafter, the new data owner is considered for anonymization and GQI communication in the network.

4.4.2. Dynamic leave

Data owners may leave the network due to unforeseen situations like power failure, system failure, and network failure. In such situations, a data owner leaves the network dynamically. It should be handled efficiently without breaching privacy. Each equivalent class consists of k or more data owners based on the privacy requirement. When a data owner departs the network, the corresponding equivalent class will be updated as per the number of remaining data owners to maintain the k-value for privacy. After the dynamic leave if the number of data owners is less than k, then the members of the equivalent group should be released to form a new group; otherwise, privacy would be breached. If a dynamic leave does not affect the minimum k-value of the group, then no specific handling is required as it is still within the privacy policy. But if the data owner who left is the first or second leader, then the leader election process should be carried out to elect new leaders.

Dismantling an existing equivalent group to form new groups during a dynamic leave is a heavy computational process. In the proposed protocol, such situations are handled by enforcing a threshold time limit. Dynamic leave of a data owner may be temporary or permanent. In temporary leave, the data owner rejoins the network within a particular time period, whereas, in permanent leave, the data owner will not join the network for further process. Hence, the threshold time is enforced to wait for any temporary leave data owner to rejoin. This reduces the computation cost as there is no further process required. If a data owner cannot rejoin within the time limit, then the members of the group will be released and a new group is formed based on the available data owners by satisfying the k-value and new leaders are elected. Thus, the dynamic leave of a data owner is efficiently handled in the protocol without a privacy breach.

5. Experiments

We evaluate our protocol in terms of computational complexity with respect to CSA elimination. In our privacy-preserving data collection protocol, we evaluate the computational complexity of the data collection phase only. The initialization and leader election phase has a complexity similar to traditional centralized anonymization techniques. Hence, the performance of the proposed protocol can be evaluated through CSA elimination of the data collection phase.

5.1. Experimental settings

The algorithms are implemented in Python programming and executed on Quad-Core Intel i7 at 2.2 GHz with 16 GB of RAM running Mac OS 10.15.3. We experimented our protocol on real-world public available datasets: the adult (63) and the informs (64).

5.2. Experimental analysis

The efficiency of the protocol in real-world datasets is analyzed in this section. First, the analysis is done with the adult dataset. There are 32,561 records with 14 attributes available in the adult dataset. The attributes "salary" and "occupation" are considered sensitive attributes. The sensitive attributes are merged as a single attribute "occupation-salary"; thus we increased the number of sensitive attributes to 30. It should be noted that our protocol does not consider multiple sensitive attributes. The computational complexity of the protocol is evaluated with the number of sensitive attributes (s) vs. time taken (in ms) by the protocol to eliminate the CSA. Figure 2 shows the computational complexity of the adult dataset with s as the x-axis and computational complexity (ms) as the y-axis. It is observed from the graph that the computational complexity increases with the number of sensitive attributes the protocol has to deal with is increased. Since the model deals with fewer sensitive attributes, the overhead seems to be stable with a slight increase in the s value.

The informs demographic dataset consists of 102,581 records and has 18 attributes. We consider "income" as the sensitive attribute and the domain size is 23,784. Figure 3 illustrates the computational complexity of the informs dataset. It is observed that the counterfeit elimination with larger domains incurs more overhead to the protocol. In the graph, the computational

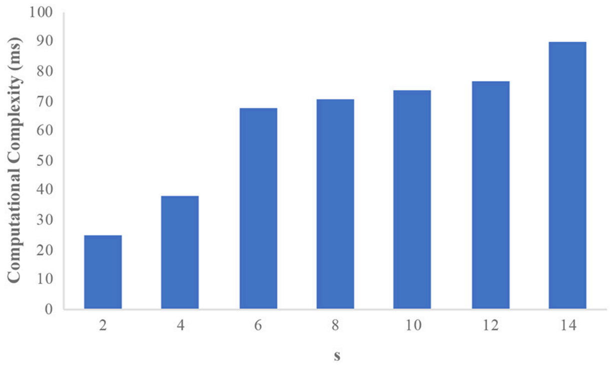

FIGURE 2
Computational complexity on the adult dataset.

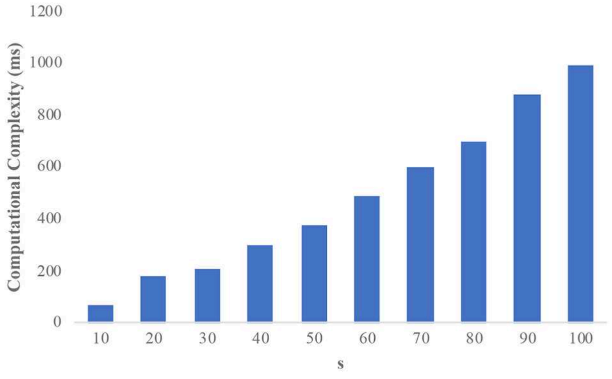

FIGURE 3
Computational complexity on the informs dataset.

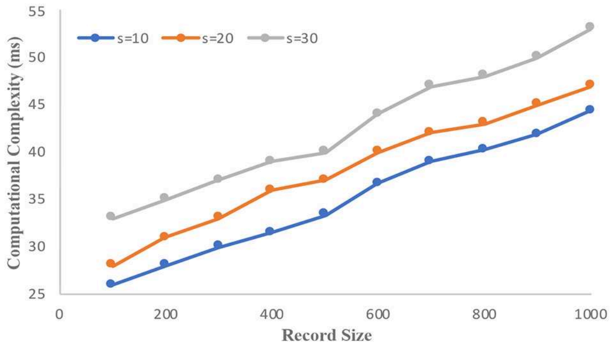

FIGURE 4
Computational complexity vs. record size.

complexity constantly increases with the size of the sensitive attributes (s) in the network. Figure 4 illustrates how the informs dataset's computing complexity varies depending on the number of sensitive features. It is understood from the graphs that computational overhead increases with the size of the dataset and the domain size. The rise is caused by the volume of fake sensitive qualities that must be addressed.

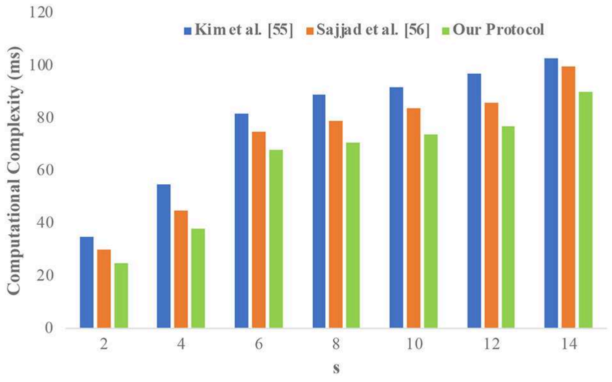

FIGURE 5
Performance evaluation of the proposed system.

5.3. Performance evaluation

The performance evaluation of the proposed study is compared with similar studies conducted by Kim and Chung (65) and Sajjad et al. (48). Figure 5 compares the performance of the proposed protocol with the state-of-the-art literature (experiments on the adult dataset). It is observed that the proposed protocol has considerably minimized the computational complexity. It is due to the slight changes in the CSA elimination where the distinct rows are compared instead of the whole dataset.

6. Discussion

This section outlines potential attacks on the suggested protocol as well as the measures the protocol uses to defend against them. We also discuss other important issues in the protocol such as leader collision mitigation and determination of CSA count. Furthermore, we discuss the complexity analysis and data utility of the protocol.

6.1. Internal and external identity disclosure attacks

When a legitimate member in the anonymization network tries to determine a person's identity, internal identity disclosure occurs. In our protocol, we consider a data collector as an adversary who seeks to gain more information about an individual. The adversary may target an individual to discern the sensitive attribute and to try to distinguish through the combination of quasi attributes. We employ a clustering-based *k*-anonymity (40) privacy model to anonymize personal health records that prevent identity disclosure. Clustering-based *k*-anonymity model generalizes the quasi attributes and forms clusters that contain at least *k* records each. As a result, the probability of identity disclosure is limited to *1/k* or less. Although the adversarial data collector has access to the generalized quasi attributes and sensitive attributes, the clustering-based *k*-anonymity policy makes internal identity disclosure nearly impossible.

External identity disclosure can happen when the data is transmitted using the given network. A practical data transmission environment is considered in the protocol, so it is necessary to add headers to the microdata. Our proposed protocol avoids direct connection between the data owner and collector in order to protect the external identity exposure, and instead relies on representatives (such as group leaders) to deliver the data to the data collector. Since all data owners in an equivalent group share the same generalized quasi attributes and the sensitive attributes are covered by a list of CSA, the group leaders are unable to determine who the data owners are. In addition to the original sensitive property, every record in the first leader dataset also contains at least *k-1* CSA. This ensures that the representative's identity disclosure does not exceed *1/k*.

6.2. Leader collision mitigation

Leader collision is a privacy attack where the elected representatives are adversarial data owners and attempt to discern sensitive information. In the leader election phase, each equivalent group elects two leaders. The first leader gathers the sensitive attribute along with the CSA. The second leader collects the CSA without real sensitive attributes and QIs. In an equivalent group if a single data owner is elected as the first and second leader, then the sensitive attributes can be discerned through the elimination of second leader sensitive attributes from the first leader dataset. In the proposed protocol, we verify the elected leaders' userIDs to make sure they are of a single data owner. Algorithm 3 shows the steps to elect different data owners as representatives.

Another type of LCA is identified by Sajjad et al. (48). Adversarial data owners may join the network by generating fake quasi attributes. They intend to be grouped under a particular equivalent group and try their chance to be elected as the group leaders. If both first and second leaders are elected from the adversarial data owners, they can collaborate and discern the sensitive attribute. This type of attack is called LCA. In our proposed protocol, we utilized firefly with a cyclic randomization algorithm (62) to elect the leaders. First, the number of data owners and their userIDs (index values) are collected, and based on the minimum and maximum index values, the random function generates two different userIDs. The generated userID is then considered the first and second leader for that specific data collection phase. The leader information is then stored in the leader information table for further verification.

6.3. Determination of counterfeit sensitive attribute count

Counterfeit sensitive attributes play an important role in protecting the sensitive attributes of the equivalent group. Similar privacy preserving data collection studies (48, 65) proposed the method of adding CSA to the anonymization network. However, the number of CSA to be added to the original sensitive attribute is not specified. It is important to determine the number of CSA required to protect the sensitive attribute in the anonymization network. In our protocol, we determine the count of CSA based on the privacy parameter k. It is proved from the k-anonymity-based privacy model that the identity can be disclosed only at the probability of $1/k$. So, we consider the privacy parameter k as the count of CSA along with the actual sensitive attribute. Hence, the sensitive attribute of each data owner is protected and the probability to disclose the sensitive attribute is not $>1/k$. In our protocol, the privacy parameter value k is shared with every data owner as the CSA count. Each data owner generates k-1 counterfeit attributes to be added with the real sensitive attributes. To improve the quality of CSA, semantic diversity (66) among the sensitive attributes can be pitched in.

6.4. Complexity analysis

The complexity of the proposed protocol can be analyzed for the three phases of the data collection protocol: initialization, leader election, and data collection phase. The data owner's initialization phase comprises *QI* generation, submission, and *GQI* validation tasks. Let Ct_{gen}, Ct_{sub}, *and* Ct_{val} be the complexity of the three tasks. *QI* generation is the basic operation of the data owner, the cost Ct_{gen} is in $O\,(1)$ where the *QI* is generated at a constant time. The complexity of Ct_{sub} is in $O\,(1)$ where each data owner can submit the *QI* at a given time. Ct_{val} is in $O\left(k\right)$ where k is the number of records in each equivalent group. In the data collector's initialization phase, the major tasks are *QI* generalization and *QI* distribution. Let Ct_{anon} and Ct_{dist} be the cost of the two tasks. Ct_{anon} is the cost of the anonymization technique that is adopted in the protocol. In traditional k-anonymity models, the cost of anonymization is NP-hard with complexity $O(n^2)$. In our proposed protocol, we adopted a clustering-based k-anonymity model so the cost Ct_{anon} is in $O\left(\frac{n^2}{k}\right)$. The distribution cost Ct_{dist} is in $O(n)$ where n is the number of data owners in the network. The total cost of the data collector at the initialization stage is in $O\left(\frac{n^2}{k}\right) + O\,(n)$. Leader election is another trivial task, the cost of Ct_{elec} is in $O\,(u)$, where u denotes the users in the network. In an equivalent group, Ct_{elec} is in $O\left(k\right)$, where k is the records in the equivalent class.

In the data collection phase, the elimination of CSA from the first leader dataset using the dataset of the second leader is a major task. The CSA values obtained from the second leader dataset are required to be compared with anonymized records of the first leader dataset. Let s be the sensitive attributes in an equivalent group then the number of sensitive attributes in a group is $k \times s$. The list of CSA in the dataset is $k \times s - 1$. If g is the number of equivalent classes, then the cost of CSA elimination is $O(g \cdot k^2 \cdot s^2)$. In our protocol, counterfeit elimination is carried out by comparing the CSA only with distinct sensitive attributes. Hence, the cost of CSA elimination is restricted to $O(kds)$ where d denotes the sensitive attribute domain size.

6.5. Data utility

In the process of anonymization, the original dataset tends to suffer from poor data utility. The data utility is generally measured through various information loss metrics. Likewise, a dataset with minimum or no information loss may leak privacy. Hence, it is important to maintain the trade-off between privacy and data utility. In our protocol, the anonymization process is carried out only during the initialization phase. The QI attributes are anonymized by the data collector through a utilized clustering-based *k*-anonymity model (53) that forms clusters as the equivalent groups with *k* or more records in each group. Thus, data utility is inherited from the adopted privacy model. Furthermore, our protocol can adopt any *k*-anonymity based privacy model. The information loss and data utility are based on the chosen privacy model. Hence, in this study, we did not present the results of the information loss as our protocol is independent of the privacy model.

6.6. Healthcare data security analysis

Beyond privacy protection, it is also essential to secure healthcare data from unauthorized access and disclosure (67). The potential security threats to a healthcare system are covered in this section.

Due to the requirements of the legal, ethical, and medical domains, healthcare data must be protected from unauthorized access and disclosure (68). To protect health information, three data security techniques are widely in use; they are cryptographic security, blockchain based security, and network security. Cryptography is the most commonly used technique to protect data from unauthorized access, tampering, and an interception. Data encryption plays a major role in protecting data. Qiu et al. (69) proposed a selective encryption algorithm to secure healthcare data sharing with fragmentation and dispersion techniques. This algorithm ensures data safety even when the cloud servers and keys are compromised. Blockchain based security techniques are popular because of their unhackable distributed ledger and smart contracts. Zhuang et al. (70) proposed a blockchain model to protect patient records from unauthorized access and disclosure. The blockchain properties such as immutability, smart contract, and distributed ledgers ensure data consistency, quick access, and patient authorization. The network is another essential part of the healthcare domain that needs proper security to avoid eavesdropping, intrusion, and tampering attacks. Most healthcare systems employ IoT, wireless networks, and body area networks. So appropriate network security is required to protect the data transferred between the data owner and the collector (71–73).

7. Conclusion and future work

In this article, we presented a privacy-preserving healthcare data collection protocol. The state-of-the-art privacy-preserving data collection models, coerce strict assumptions such as secure private channels or third-party anonymization between the data owners and the collector. The proposed protocol eliminates such assumptions and offers anonymous data collection through the elected representatives among the data owners. The protocol is efficient in tackling internal and external identity disclosure through an adopted clustering-based *k*-anonymity model. We proposed solutions for possible collisions among the elected representatives within the equivalent group. We also proposed a new efficient method to add CSA to protect the real sensitive attributes. Furthermore, dynamic data owners are efficiently handled in the protocol by a greedy heuristic method. Through extensive experimental analysis, we proved that our protocol incurs considerably minimum computational complexity compared with state-of-the-art techniques. This makes our protocol more suitable for collecting huge amounts of healthcare datasets without privacy breach. Our protocol is built to accommodate any *k*-anonymity-based privacy models; hence, the data utility can be optimized as per the requirement.

We intend to conduct several future studies to address the limitations of this study. First, we would like to focus on minimizing the other privacy risks such as attribute disclosure, membership disclosure, and similarity attacks. Currently, our study is focused mainly on protecting personal data from identity disclosure. Considering other privacy attacks would make our protocol more robust for healthcare data collection. Second, we would like to employ anonymization techniques other than *k*-anonymity such as bucketization and anatomy to enhance the data utility of the protocol.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://archive.ics.uci.edu/ml/datasets/adult.

Author contributions

JA and JK conceived the idea and worked on the technical details. JA, RE, and JK devised the work, the main conceptual ideas, the proof outline, and worked on the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Varshney U. Pervasive healthcare and wireless health monitoring. *Mob Networks Appl.* (2007) 12:113–27. doi: 10.1007/s11036-007-0017-1

2. Andreassen HK, Bujnowska-Fedak MM, Chronaki CE, Dumitru RC, Pudule I, Santana S, et al. European citizens' use of E-health services: a study of seven countries. *BMC Public Health.* (2007) 7:1–7. doi: 10.1186/1471-2458-7-53

3. Benharref A, Serhani MA. Novel cloud and SOA-based framework for E-health monitoring using wireless biosensors. *IEEE J Biomed Heal Informat.* (2014) 18:46–55. doi: 10.1109/JBHI.2013.2262659

4. Aldosari B. Patients' safety in the era of EMR/EHR automation. *Informatics Med Unlocked.* (2017) 9:230–3. doi: 10.1016/j.imu.2017.10.001

5. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform.* (2008) 77:291–304. doi: 10.1016/j.ijmedinf.2007.09.001

6. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet.* (2012) 13:395–405. doi: 10.1038/nrg3208

7. Saleh S, Abdouni L, Dimassi H, Nabulsi D, Harb R, Jammoul Z, et al. Prevalence of non-communicable diseases and associated medication use among Syrian refugees in Lebanon: an analysis of country-wide data from the Sijilli electronic health records database. *Confl Health.* (2021) 15:77. doi: 10.21203/rs.3.rs-58159/v1

8. Saleh S, Alameddine M, Farah A, Arnaout NE, Dimassi H, Muntaner C, et al. eHealth as a facilitator of equitable access to primary healthcare: the case of caring for non-communicable diseases in rural and refugee settings in Lebanon. *Int Public Health J.* (2018) 63:577–88. doi: 10.1007/s00038-018-1092-8

9. Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, et al. Use of Electronic Health Records in U.S. Hospitals. *N. Engl. J. Med.* (2009) 360:1628–38. doi: 10.1056/NEJMsa0900592

10. Blumenthal D, Tavenner M. The 'meaningful use' regulation for electronic health records. *N Engl J Med.* (2010) 363:501–4. doi: 10.1056/NEJMp1006114

11. Garde S, Knaup P, Hovenga EJS, Heard S. Towards semantic interoperability for electronic health records: domain knowledge governance for openEHR archetypes. *Methods Inf Med.* (2007) 46:332–43. doi: 10.1160/ME5001

12. Lingren T, Sadhasivam S, Zhang X, Marsolo K. Electronic medical records as a replacement for prospective research data collection in postoperative pain and opioid response studies. *Int J Med Inform.* (2018) 111:45–50. doi: 10.1016/j.ijmedinf.2017.12.014

13. Haas S, Wohlgemuth S, Echizen I, Sonehara N, Müller G. Aspects of privacy for electronic health records. *Int J Med Inform.* (2011) 80:e26–31. doi: 10.1016/j.ijmedinf.2010.10.001

14. Demuynck L, De Decker B. Privacy-preserving electronic health records. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3677* (Salzburg: Springer Berlin Heidelberg). (2005). p. 150–9.

15. Rizwan M, Shabbir A, Javed AR, Srivastava G, Gadekallu TR, Shabir M, et al. Risk monitoring strategy for confidentiality of healthcare information. *Comput Electr Eng.* (2022) 100:107833. doi: 10.1016/j.compeleceng.2022.107833

16. El Zarif O, Haraty RA. Toward information preservation in healthcare systems. *Innov Heal Informat A Smart Healthc Prim.* (2020) 163–85. doi: 10.1016/B978-0-12-819043-2.00007-1

17. Xue M, Papadimitriou P, Raïssi C, Kalnis P, Pung HK. Distributed privacy preserving data collection. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 6587* (Hong Kong: Springer Berlin Heidelberg). (2011). p. 93–107.

18. Andrew J, Karthikeyan J, Jebastin J. Privacy preserving big data publication on cloud using mondrian anonymization techniques and deep neural networks. In: *2019 5th International Conference on Advanced Computing and Communication Systems.* (2019). p. 722–7.

19. Dhasarathan C, Hasan MK, Islam S, Abdullah S, Mokhtar UA, Javed AR, et al. COVID-19 health data analysis and personal data preserving: a homomorphic privacy enforcement approach. *Comput Commun.* (2023) 199:87–97. doi: 10.1016/j.comcom.2022.12.004

20. Haraty RA, Boukhari B, Kaddoura S. An effective hash-based assessment and recovery algorithm for healthcare systems. *Arab J Sci Eng.* (2022) 47:1523–36. doi: 10.1007/s13369-021-06009-4

21. Liu YN, Wang YP, Wang XF, Xia Z, Xu JF. Privacy-preserving raw data collection without a trusted authority for IoT. *Comput Networks.* (2019) 148:340–8. doi: 10.1016/j.comnet.2018.11.028

22. Sei Y, and Okumura AJH, Ohsuga A. Privacy-preserving collaborative data collection and analysis with many missing values. *IEEE Trans. Dependable Secur. Comput.* (2022). doi: 10.1109/TDSC.2022.3174887

23. Krasnova H, Günther O, Spiekermann S, Koroleva K. Privacy concerns and identity in online social networks. *Identity Inf Soc.* (2009) 2:39–63. doi: 10.1007/s12394-009-0019-1

24. Fung BCM, Wang K, Yu PS. Anonymizing classification data for privacy preservation. *IEEE Trans Knowl Data Eng.* (2007) 19:711–25. doi: 10.1109/TKDE.2007.1015

25. Byun JW, Kamra A, Bertino E, Li N. Efficient k-anonymization using clustering techniques. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 4443* (Bangkok: Springer Berlin Heidelberg). (2007). p. 188–200.

26. Zakerzadeh H, Osborn SLAANST. Fast anonymizing algorithm for numerical streaming data. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 6514.* (2011). p. 36–50.

27. Prakash M, Singaravel G. An approach for prevention of privacy breach and information leakage in sensitive data mining. *Comput Electr Eng.* (2015) 45:134–40. doi: 10.1016/j.compeleceng.2015.01.016

28. Li HT, Ma JF, Fu SA. privacy-preserving data collection model for digital community. *Sci China Inf Sci.* (2015) 58:1–16. doi: 10.1007/s11432-014-5197-2

29. Yang Z, Zhong S, Wright RN. Anonymity-preserving data collection. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* (2005). p. 334–43.

30. Erlingsson Ú, Pihur V, Korolova A. Rappor: randomized aggregatable privacy-preserving ordinal response. In: *Proceedings of the ACM Conference on Computer and Communications Security.* (2014). p. 1054–67.

31. Kim JW, Jang B, Yoo H. Privacy-preserving aggregation of personal health data streams. *PLoS ONE.* (2018) 13:e0207639. doi: 10.1371/journal.pone.0207639

32. Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: a survey of recent developments. *ACM Comput. Surv.* (2010) 42. doi: 10.1145/1749603.1749605

33. Wang M, Xiao D, Liang J, Hu G. Distributed privacy-preserving nested compressed sensing for multiclass data collection with identity authentication. *Signal Process.* (2023) 204:108823. doi: 10.1016/j.sigpro.2022.108823

34. Zhang W, Jiao H, Yan Z, Wang X, Khan MK. Security analysis and improvement of a public auditing scheme for secure data storage in fog-to-cloud computing. *Comput Secur.* (2023) 125:103019. doi: 10.1016/j.cose.2022.103019

35. Kaaniche N, Laurent M. Data security and privacy preservation in cloud storage environments based on cryptographic mechanisms. *Comp Commun.* (2017) 111:120–41. doi: 10.1016/j.comcom.2017.07.006

36. Puri V, Sachdeva S, Kaur P. Privacy preserving publication of relational and transaction data: survey on the anonymization of patient data? *Comp Sci Rev.* (2019) 32:45–61. doi: 10.1016/j.cosrev.2019.02.001

37. Sweeney L. k-anonymity: a model for protecting privacy. *Int Uncert J Fuzziness Knowl Based Syst.* (2002) 10:557–70. doi: 10.1142/S0218488502001648

38. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. ℓ-Diversity: privacy beyond k-anonymity. *Proc Int Conf Data Eng.* (2006) 2006:24. doi: 10.1109/ICDE.2006.1

39. Ninghui L, Tiancheng L, Venkatasubramanian S. t-closeness: privacy beyond k-anonymity and ℓ-diversity. in *Proceedings - International Conference on Data Engineering.* (2007). p. 106–15.

40. Lin JL, Wei MC. An efficient clustering method for k-anonymization. *ACM Int Conf Proc Ser.* (2008) 331:46–50. doi: 10.1145/1379287.1379297

41. Li H, Guo F, Zhang W, Wang J, Xing J. (a,k)-anonymous scheme for privacy-preserving data collection in IoT-based healthcare services systems. *J Med Syst.* (2018) 42:1–9. doi: 10.1007/s10916-018-0896-7

42. Truta TM, Vinay B. Privacy protection: P-sensitive k-anonymity property. In: *ICDEW 2006 - Proceedings of the 22nd International Conference on Data Engineering Workshops* (2006).

43. Zhong S, Yang Z, Wright RN. Privacy-enhancing k-anonymization of customer data. In: *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.* (2005). p. 139–47.

44. Zhong S, Yang Z, Chen T. k-anonymous data collection. *Inf Sci.* (2009) 179:2948–63. doi: 10.1016/j.ins.2009.05.004

45. Hussien AA, Hamza N, Hefny HA. Attacks on anonymization-based privacy-preserving: a survey for data mining and data publishing. *J Inf Secur.* (2013) 04:101–12. doi: 10.4236/jis.2013.42012

46. Beg S, Anjum A, Ahmad M, Hussain S, Ahmad G, Khan S, et al. A privacy-preserving protocol for continuous and dynamic data collection in IoT enabled mobile app recommendation system (MARS). *J Netw Comput Appl.* (2021) 174:102874. doi: 10.1016/j.jnca.2020.102874

47. Beg S, Anjum A, Ahmed M, Malik SUR, Malik H, Sharma N, et al. Dynamic parameters-based reversible data transform (RDT) algorithm in recommendation system. *IEEE Access.* (2021) 9:110011–25. doi: 10.1109/ACCESS.2021.3101150

48. Sajjad H, Kanwal T, Anjum A, Khan A, Khan A, Manzoor U. An efficient privacy preserving protocol for dynamic continuous data collection. *Comput Secur.* (2019) 86:358–71. doi: 10.1016/j.cose.2019.06.017

49. Ni C, Cang LS, Gope P, Min G. Data anonymization evaluation for big data and IoT environment. *Inf Sci.* (2022) 605:381–92. doi: 10.1016/j.ins.2022.05.040

50. Sei Y, Andrew Onesimu J, Ohsuga A. Machine learning model generation with copula-based synthetic dataset for local differentially private numerical data. *IEEE Access.* (2022) 1:101656–101671. doi: 10.1109/ACCESS.2022.3208715

51. Sun S, Ma S, Song JH, Yue WH, Lin XL, Ma T, et al. Experiments and analyses of anonymization mechanisms for trajectory data publishing. *J Comput Sci Technol.* (2022) 375:1026–48. doi: 10.1007/s11390-022-2409-x

52. Canbay Y, Sagiroglu S, Vural YA. new utility-aware anonymization model for privacy preserving data publishing. *Concurr Comput Pract Exp.* (2022) 34:e6808. doi: 10.1002/cpe.6808

53. Onesimu JA, Karthikeyan J, Sei Y. An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services. *Peer Peer Netw Appl.* (2021) 14:1629–49. doi: 10.1007/s12083-021-01077-7

54. Abbasi A, Mohammadi BA. Clustering-based anonymization approach for privacy-preserving in the healthcare cloud. *Concurr Comput Pract Exp.* (2022) 34:e6487. doi: 10.1002/cpe.6487

55. Onesimu JA, Karthikeyan J, Eunice J, Pomplun M, Dang H. Privacy preserving attribute-focused anonymization scheme for healthcare data publishing. *IEEE Access.* (2022) 10:86979–97. doi: 10.1109/ACCESS.2022.3199433

56. Avraam D, Jones E, Burton PA. deterministic approach for protecting privacy in sensitive personal data. *BMC Med Inform Decis Mak.* (2022) 22:1–17. doi: 10.1186/s12911-022-01754-4

57. Kanwal T, Anjum A, Khan A, Asheralieva A, Jeon GA. Formal adversarial perspective: Secure and efficient electronic health records collection scheme for multi-records datasets. *Trans Emerg Telecommun Technol.* (2021) 32:e4180. doi: 10.1002/ett.4180

58. Kanwal T, Anjum A, Malik SU, Sajjad H, Khan A, Manzoor U, et al. A robust privacy preserving approach for electronic health records using multiple dataset with multiple sensitive attributes. *Comput Secur.* (2021) 105:102224. doi: 10.1016/j.cose.2021.102224

59. Jayapradha J, Prakash M. f-Slip: an efficient privacy-preserving data publishing framework for 1:M microdata with multiple sensitive attributes. *Soft Comput.* (2022) 26:13019–36. doi: 10.1007/s00500-021-06275-2

60. Khan SI, Khan ABA, Hoque ASML. Privacy preserved incremental record linkage. *Big Data J.* (2022) 9:1–27. doi: 10.1186/s40537-022-00655-7

61. Raju NVSL, Naresh VS. Dynamic distributed KCi-slice data publishing model with multiple sensitive attributes. *Concurr Comput Pract Exp.* (2022) 34:e7064. doi: 10.1002/cpe.7064

62. Sarkar A, Senthil Murugan T. Cluster head selection for energy efficient and delay-less routing in wireless sensor network. *Wirel Networks.* (2019) 25:303–20. doi: 10.1007/s11276-017-1558-2

63. Dua D, Graff, C. *UCI Machine Learning Repository: Adult Data Set*. UCI (2017). Available online at: https://archive.ics.uci.edu/ml/datasets/adult (accessed March 2, 2019).

64. *Data – Informsdataminingcontest.* Available online at: https://sites.google.com/site/informsdataminingcontest/data (accessed July 12, 2020).

65. Kim S, Chung YD. An anonymization protocol for continuous and dynamic privacy-preserving data collection. *Futur Gener Comput Syst.* (2019) 93:1065–73. doi: 10.1016/j.future.2017.09.009

66. Oishi K, Sei Y, Tahara Y, Ohsuga A. Semantic diversity: privacy considering distance between values of sensitive attribute. *Comput Secur.* (2020) 94:101823. doi: 10.1016/j.cose.2020.101823

67. Kondepogu MD, Andrew J. Secure E-health record sharing using blockchain: a comparative analysis study. In: *Proc - 2022 6th Int Conf Intell Comput Control Syst ICICCS 2022.* (2022). 861–8.

68. Thapa C, Camtepe S. Precision health data: requirements, challenges and existing techniques for data security and privacy. *Comp Biol Med.* (2021) 129:104130. doi: 10.1016/j.compbiomed.2020.104130

69. Qiu H, Qiu M, Liu M, Memmi G. Secure health data sharing for medical cyber-physical systems for the Healthcare 4.0. *IEEE J Biomed Heal Inf.* (2020) 24:2499–505. doi: 10.1109/JBHI.2020.2973467

70. Zhuang Y, Sheets LR, Chen YW, Shae ZY, Tsai JJP, Shyu CRA, et al. Patient-centric health information exchange framework using blockchain technology. *IEEE J Biomed Heal Informatics.* (2020) 24:2169–76. doi: 10.1109/JBHI.2020.2993072

71. Huang H, Gong T, Ye N, Wang R, Dou Y. Private and secured medical data transmission and analysis for wireless sensing healthcare system. *IEEE Trans Ind Informatics.* (2017) 13:1227–37. doi: 10.1109/TII.2017.2687618

72. Zhang Y, Deng RH, Han G, Zheng D. Secure smart health with privacy-aware aggregate authentication and access control in internet of things. *J Netw Comput Appl.* (2018) 123:89–100. doi: 10.1016/j.jnca.2018.09.005

73. Andrew J, Kathrine GJW. An intrusion detection system using correlation, prioritization and clustering techniques to mitigate false alerts. *Adv Big Data Cloud Comp*. (2018) 645:257–68. doi: 10.1007/978-981-10-7200-0_23