Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2012, Article ID 521267, 9 pages
doi:10.1155/2012/521267

Research Article

A Privacy-Preserved Analytical Method for eHealth Database with Minimized Information Loss

Ya-Ling Chen,[1] Bo-Chao Cheng,[2] Hsueh-Lin Chen,[1] Chia-I Lin,[1] Guo-Tan Liao,[2] Bo-Yu Hou,[2] and Shih-Chun Hsu[2]

[1] *Service Systems Technology Center, Industrial Technology Research Institute (ITRI), Hsinchu 31040, Taiwan*
[2] *Department of Communications Engineering, National Chung Cheng University, Chiayi 62145, Taiwan*

Correspondence should be addressed to Guo-Tan Liao, loboyoh@gmail.com

Received 17 May 2012; Revised 18 July 2012; Accepted 19 July 2012

Academic Editor: Tai Hoon Kim

Digitizing medical information is an emerging trend that employs information and communication technology (ICT) to manage health records, diagnostic reports, and other medical data more effectively, in order to improve the overall quality of medical services. However, medical information is highly confidential and involves private information, even legitimate access to data raises privacy concerns. Medical records provide health information on an as-needed basis for diagnosis and treatment, and the information is also important for medical research and other health management applications. Traditional privacy risk management systems have focused on reducing reidentification risk, and they do not consider information loss. In addition, such systems cannot identify and isolate data that carries high risk of privacy violations. This paper proposes the Hiatus Tailor (HT) system, which ensures low re-identification risk for medical records, while providing more authenticated information to database users and identifying high-risk data in the database for better system management. The experimental results demonstrate that the HT system achieves much lower information loss than traditional risk management methods, with the same risk of re-identification.

1. Introduction

Electronic medical records and cloud storage have been introduced in hospitals in recent years. Medical institutions are required to store electronic records in a database and provide access for doctors and researchers. Digital records [1, 2] provide convenience, but such a system also introduces the new challenge of storing personal information securely. The issue of privacy [3] has received much public attention recently. Based on personal information, a specific person can be identified directly or indirectly. Information that can be used to directly identify a particular person is called personally identifiable information (PII). According to the definition given by the United States Office of Management and Budget, full name, Social Security Number, face, fingerprints, and genetic information are all categorized as PII.

According to NIST IR7628, personal information privacy means a person has the right to decide when and where to disclose their personal information. It also says that the storage and access of personal information and PII must be secure. Three personal information security measures have been proposed in NIST SP800-122: (1) minimizing the use, collection, and retention of PII, (2) conducting privacy impact assessments, and (3) deidentifying information.

Medical institutions save large amounts of personal information in databases whose contents can be divided into three categories: Direct Identifiers (DID), Quasi-identifiers (QID), and Sensitive Information (SI). Information that allows direct identification, such as the Social Security Number, is called DID. Details such as date of birth, level of education, and postcode, which can be combined to identify a person, are QID. Information that is private and confidential, such as medical conditions, is categorized as

SI. To provide security of personal information, medical institutions are required to check information before release to prevent any violation of patient privacy.

When eHealth practitioners (such as service provider, insurance company and other health researcher) want to access medical records, the hospital can de-identify the database to protect patient privacy. However, when multiple users need to access the database, they would each have unique requirements. The hospital must release several de-identified databases, which are then difficult to manage. In addition, the de-identified database differs from the original database. In other words, the de-identified database will be altered and the degree of alteration is represented by the information loss (IL). As the database provider, the hospital prefers high IL to protect patient privacy and lower the possibility of re-identification of the information. In contrast, researchers prefer databases with low IL for their work. Therefore, the challenge is to strike a balance between the two interests.

An information management procedure has been proposed [4] to manage research-oriented electronic medical records. The aim is to minimize the probability of disclosure of personal information. The procedure is as follows.

(1) The information owner must check the legitimacy of the reason for requiring access to the database.

(2) A risk assessment must be conducted based on the user's requirements.

(3) Decide whether de-identification is needed based on the risk associated. Execute various de-identification methods.

(4) Release the database to a user once the risk of re-identification is acceptable.

De-identification [5, 6] is the primary method of protecting private information, where the original database is modified to prevent direct identification of a person through their records even if multiple databases are combined. Some common de-identification techniques are data reduction, data modification, data suppression, perturbation, and pseudonymisation [7]. The k-anonymity model [8–10] is commonly used to assess the performance of a de-identification technique in reducing the risk of re-identification. When users search the database after a database is de-identified, one of every k results is authentic. However, the other $k - 1$ results also appear in the search results. Usually, the authenticity of the results cannot be determined, which means the higher the k value is, the lower the risk of re-identification is [11].

Currently, numerous privacy-preserving administration tools are commercially available on the market, five of which are markedly popular [12]: the PARAT, μ-Argus, CAT, UTD Toolbox, and sdMicro. Among them, the UTD Toolbox and CAT are based on the k-anonymity algorithm. The UTD Toolbox does not provide active support for its products, despite its functions designed from the developer's perspective. The CAT suffers from usability difficulties. For example, because the k value of k-anonymity cannot be defined using the CAT, this tool operates unstably. In contrast to the CAT, the sdMicro is unable to process large datasets; furthermore, it crashes frequently. Currently, the tool receiving the most support is the PARAT, which is superior to CAT regarding the k-anonymity algorithm, and outperforms the μ-Argus in resulting precision level.

Some previous studies have focused on reducing the risk of re-identification. However, limited research effort has been spent on safeguarding privacy while minimizing data distortion. El Emam et al. [13] proposed a set of programs that balance the risk and the extent of data distortion. If the risk exceeds the preset threshold value, the system tests various de-identification techniques to try and limit data distortion to the required level. However, such a system is unable to identify the data that is responsible for the higher risk effectively; it spends a lot of time on the trial-and-error process.

In this study, we propose the Hiatus Tailor (HT) system. By using the Execution Chain Graph (ECG) to progressively de-identify data, people's privacy can be protected. The name Hiatus Tailor refers to the fact that the proposed system is capable of identifying the missing element within the system and fixing it. It uses progressive risk assessment and mitigation, and is able to balance the risk of re-identification and data distortion. Among the scenarios where the re-identification risk requirement is satisfied, the proposed method chooses the one that minimizes the distortion level. The main contributions of this paper are summarized as follows.

(i) In contrast to other de-identification methods that de-identify the entire database once, resulting in high IL, the HT system not only meets the privacy protection requirements, but also categorizes data into QID blocks using ECG. The risk is assessed progressively for each block. Based the re-identification risk estimated by this assessment, an optimal de-identification method is selected. As de-identification is not required at every node, the HT system is capable of reducing IL.

(ii) Tradition risk assessment methods can only indicate whether the risk is high or low. However, for most databases, the source of the risk cannot be identified. Therefore, the process of identifying the source of the increased risk is time consuming. The HT system uses QID and progressively assesses risk for a database. ECG allows an examination of the entire system and assists medical institutions in evaluating whether the target system satisfies privacy safeguard requirements. If the system is found to have a high level of risk, it is easier to identify and handle the QID data block that is responsible for the high-risk level.

2. HT System Architecture and Operation Method

The two main components of the HT system architecture are the Execution Chain Graph Composer (ECG Composer) and the Privacy Tailor. Based on various user requirements, the ECG composer creates the Execution Chain Graph and

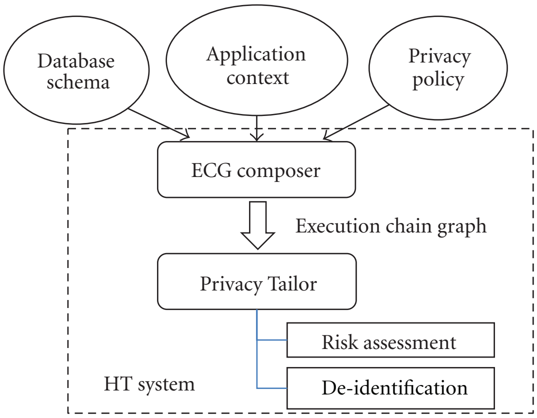


Figure 1: HT system architecture.

sends it to the Privacy Tailor. As the Privacy Tailor receives the Execution Chain Graphs from the ECG Composer at different nodes of execution, it assesses the risk of QID combinations in the database. If the risk is too high, it de-identifies the identifiable information with less information loss in the database.

2.1. Architecture. The HT system architecture consists of two major components: ECG Composer and the Privacy Tailor (as shown in Figure 1). ECG Composer compiles the information obtained from users' requirements and generates the Execution Chain Graph, which is sent to the Privacy Tailor for further processing and risk assessment.

The operation of the ECG Composer is based on information from the following elements.

(i) Database schema: defines the properties of the database, such as the type of the tables in the database and the attributes of the table. From the database schema, the data types of the stored data can be identified.

(ii) Application context: includes components related to SQL query statements, which is performed with the SELECT statement to retrieve a list of columns (including QIDs and other regular data) from one or more queried tables with the optional WHERE clause only returning the rows for which the comparison predicate evaluates to True. These SQL query statements are the details relevant to the user application. The order in which the application accesses QIDs determines which QIDs are analyzed by ECG in different nodes.

(iii) Privacy policy: defines the privacy policy associated with the user or company, such as the threshold k (k-anonymity) for the QID. The privacy policy is modeled as (U, Q, K, G, and F), for different users (U), the administrator can specify the QID(Q) list, the threshold k (K) to be satisfied for k-anonymity, and the de-identification technique (G). The file (F) of the de-id technique contains the de-id policy where we adopt the taxonomy tree approach described in [14]. The de-identification technique (G) may include Data Reduction, Data Modification, Data Suppression, Pseudonymisation, and Generalization. Each de-identification technique has its own specification which is described in the file (F). For instance, Generalization technique will revise the attributes in a hierarchy manner based on the taxonomy tree structure described in file (F). Take the field "country of origin" as a Generalization technique example. USA and Canada are part of North-America. If they are generalized, both USA and Canada will be represented as North America.

Based on user requirements, ECG Composer compiles the information obtained from these components and generates the Execution Chain Graph, which is sent to the Privacy Tailor for further processing and risk assessment.

Privacy Tailor is analogous to a privacy management department. Its operation can be described as two stages: (1) risk assessment: executes the risk assessment procedure and estimates the re-identification risk of the current assessment phase. (2) Deidentification: on completing the risk assessment, if the re-identification risk is higher than the threshold, Privacy Tailor identifies the tuples that has relatively high risk and needs to be de-identified. The re-identification risk is calculated as described in [15] (as shown in (1)):

$$R = \frac{1}{\mathrm{Min}_j\left(F_j\right)}, \tag{1}$$

where F_j is the size of an equivalence class.

An equivalence class is the set of records in the database which have the same values on all quasi-identifier attributes. When an equivalence classes has the smallest value, we have the highest probability of re-identification and use it as our re-identification risk. As such, the Risk Assessment component will scan the database based on various de-identified QID combinations to find the size for each equivalence class and obtain the re-identification risk.

ECG Composer uses the contents of the Database schema provided by the user, the operations defined in Application context, and the privacy policy associated with the user, to generate a series of Execution Chain Graphs and forward them to Privacy Tailor. The Execution Chain Graph will be described in the next section. Both Privacy Tailor and Execution Chain Graph node use a node as their unit and are divided based on several levels of re-identification risk of the QID combination in the required database table. When the re-identification risk is below the privacy policy threshold, no operations are required; Privacy Tailor continues to the next node. When the re-identification value is larger than the privacy policy threshold, identification is performed at that level by comparing the re-identification risk value for different combinations of QID to find the most suitable scheme.

2.2. Execution Chain Graph (ECG). Database access task execution is modeled and structured in various stages aimed at clients in several nodes of database retrievals. As described earlier, the ECG Composer compiles the user requirements, consisting of the Database Schema, Application Context, and Privacy Policy, and then generates the Execution Chain Graph in which each node represents a “stored procedure” that accesses database system, and the directed edge denotes execution sequence (or caller to caller relations). Each stage consists of several atomic “stored procedure” nodes which have a set of associated attributes as follows.

(i) Information loss: the magnitude of the difference between the original database and the database after de-identification.

(ii) Re-identification risk: the possibility of identifying a specific entity directly or indirectly with various de-identified QID combinations.

(iii) Table access: the table name where information is stored and accessed.

(iv) QID: quasi-identifier, which is a subset of attributes that can indirectly identify a specific entity in a table.

(v) Condition: the relevant WHERE clause of the SQL statement is used to extract the records which satisfy a specified criterion.

These properties can be further classified as Local and Aggregate. The Local value is the result of evaluating the QID combination of the current node. Aggregate value is the result of adding the evaluation of all QID combinations of all previous nodes.

2.3. ECG Composer. This section describes the ECG composer process. The ECG composer requires users to provide relevant data as input. When the system receives data from the admin, it will output an Execution Chain Graph according to requirements, and each node will have a form to record relevant data. The input to ECG composer consists of the Database Schema Ω; Application Context Ψ; and QID List Γ. Algorithm 1 shows the algorithm of ECG composer, which creates a node set S based on the user's Application Context. Every node has an associated form that records node information. The order in which the application accesses QIDs determines the execution order which represents a direct edge from S_i to its successor, S_j. It will retrieve the specified table, attribute list (AL), and conditions for the data from the Application Context. ECG composer compares the AL with the QID list (QL). If there is an intersection, the QIDs in the intersection will be assessed according to the privacy policy, in the order of application access. In each node, node information will be updated to complete ECG generation.

Figure 2 shows an example for the operations of ECG composer. Supposedly, we have QID List (Γ = age, region, sex, and education) and Application Context Ψ listed as below:

SELECT age FROM E_table WHERE age ≥30,

SELECT region FROM E_table WHERE age ≥30,

SELECT sex FROM E_table WHERE age ≥30.

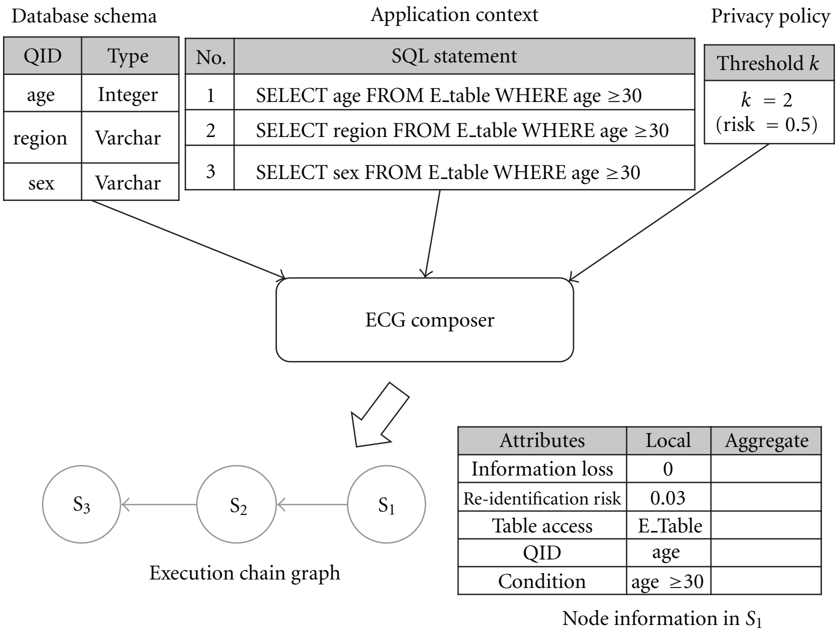
Database Schema defines the data types for age, region, and sex as integer, varchar, varchar, respectively. Based on line 5 and 6 in Algorithm 1, ECG composer creates a node set S with 3 nodes (S_1, S_2, and S_3) and connects the 3 nodes. Each node has an empty node information form that specifies information loss, re-id risk, and table access. This is the initial ECG. For each node, ECG executes line 08 statement to extract the (Table, AL, Condition) from Ψ. For example, (E_table, age, age ≥ 30) is extracted from the SQL statement “SELECT age FROM E_table WHERE age ≥ 30” for S_1 . Next, ECG composer will compute the intersection of the attribute list (e.g., AL = age for S_1) and the QID List (Γ = age, region, sex, and education). If the intersection (QL) is not empty then ECG performs two steps (line 11 and line 12) as follows: (1) updates node information form (TABLE, QL, Condition) for S_i; and (2) assesses risk for the current node S_i locally.

In our example, according to the order of application access, the system will assess age, region and sex in S_1, S_2, and S_3 one by one. The assessment is based on the threshold k defined in the input privacy policy. For example, in node S_1, according to SQL statement (SELECT age FROM E_table WHERE age ≥ 30), the age data from E_table satisfying age ≥ 30 will be selected and by the definition in database schema, age is an integer value. After risk assessment, the re-id risk is calculated to be 0.03. Initially, as the data has not been processed yet, the value of IL is 0. When node information is updated, IL = 0, re-id risk = 0.03, Table Access = E_table, QID = age, and Condition = age ≥ 30 will be recorded in the node information. On the other hand, when the intersection (QL) is empty which means this SQL statement has no risk due to no QID access, we will skip the node S_i.

2.4. Privacy Tailor. Algorithm 2 represents the Privacy Tailor algorithm. After the ECG composer creates the Execution Chain Graph, Privacy Tailor will calculate the re-identification risk and extent of data alteration at the level of the node and record it in the node data. If the risk value is higher than the threshold, Privacy Tailor will first evaluate and analyze each node to estimate re-identification risk and choose the most appropriate data for identification.

```
Given: Database schema Ω; Application Context Ψ;
    QID List Γ;
AL: Attribute list;
QL: Target QID list;
S: node set;
  //Create ECG
S = Construct the node set from Ψ;
Build the direct edge set for each pair of (S_i, S_i) based on the order of the QID accesses in Ψ;
For each node S_i{ //S_i ∈ S (i: number of node);
    Extract the corresponding (Table, AL, Condition);
    QL = ∩ AL;
    If (QL != Φ) {
        Update (TABLE, QL, Condition) for S_i;
    Assess risk for S_i;
    }
}
```

Algorithm 1: ECG composer algorithm.



Database schema

QID	Type
age	Integer
region	Varchar
sex	Varchar

Application context

No.	SQL statement
1	SELECT age FROM E_table WHERE age ≥30
2	SELECT region FROM E_table WHERE age ≥30
3	SELECT sex FROM E_table WHERE age ≥30

Privacy policy

Threshold k
k = 2 (risk = 0.5)

Attributes	Local	Aggregate
Information loss	0	
Re-identification risk	0.03	
Table access	E_Table	
QID	age	
Condition	age ≥30	

Node information in S_1

Figure 2: ECG composer operation.

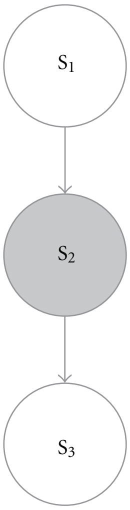
However, after knowing the identification information, the re-identification risk value will change. Therefore, the Privacy Tailor must reanalyze based the new information. If the calculated risk value does not exceed the threshold, it proceeds to the next node for analysis. When the re-identification value at each node is below the threshold, the Privacy Tailor completes execution.

Continuing the example from Figure 2, the Execution Chain Graph can be divided into three levels, node in terms of nodes S_1, S_2, and S_3 (as shown in Figure 3). Using S_1 as an example, re-identification of node information shows no value initially. Next, the Privacy Tailor performs an evaluation and fills in the current node information. In node S_1, all QIDs belong to E_table, the Age data. It satisfies the Conditions (comparison predicate) restricting the rows returned by the query (e.g., age ≥ 30), as the re-identification risk is 0.03. Thus, de-identification is no required and data distortion is zero. In addition, if risk value is larger than the user-specified threshold, the user specified de-identification method will be used and privacy model classes will be created according to the de-identification file.

Assume that a user requires access to information stored in the electronic hospital records database. The information in the database may include patients' age, region, and gender.

```
Given ECG & Threshold;
main(){
  while(ok==0){      //Set ok = 1 when finish
    recheck=0;
    Node_information=compute(ECG);
       //Compute information loss and Re-ID risk
    Re-id_risk=getRisk(Node_information);
   if(Re-id_risk > Threshold){
      target=Find_De-id_target();
      De-id(target);
     recheck=1;
     node=0;
    }
    if(recheck!=1){
      if(node_number == max){
        ok=1; //Reach the last node and end the process
      }
       else{ node_number++; }
    }
  }
} // End of main
Find_De-id_target(){
  Compute aggregate risk for each node at each node;
  Choose the highest risk node;
  return node; }
```

Algorithm 2: Privacy Tailor algorithm.



Node information in S_1

Attributes	Local	Aggregate
Information loss	0	0
Re-identification risk	0.03	0.03
Table access	E_Table	E_Table
QID	Age	Age
Condition	Age ≥30	Age ≥30

Node information in S_2

Attributes	Local	Aggregate
Information loss	0	30%
Re-identification risk	0.23	0.24
Table access	E_Table	E_Table
QID	Region	Age × region
Condition	Age ≥30	Age ≥30

Node information in S_3

Attributes	Local	Aggregate
Information loss	0	30%
Re-identification risk	0.1	0.3
Table access	E_Table	E_Table
QID	Sex	Age × region × sex
Condition	Age ≥30	Age ≥30

Figure 3: Privacy Tailor operation.

Based on the user's requirements, Privacy Tailor performs risk assessment. The detailed processes are described as follows.

(i) At node S_1 , the Privacy Tailor begins evaluation using the QID combination of the chosen table, which is the re-identification risk of the patients' age. Assuming that the threshold of the privacy policy equals to 2, the re-identification value calculated is 0.03, which is less than the threshold value 0.5. Thus, the Privacy Tailor decides that age is low risk and de-identification is not needed; the IL value is therefore 0.

(ii) After evaluating S_1, node S_2 is evaluated, which involves calculating the re-identification risk of the combination of age and region (age × region). Supposedly, the result obtained is 0.73, which exceeds the threshold. Therefore, the Privacy Tailor must proceed with de-identification at this level. There are three possible de-identification ways (age, region, and age× region), each associated with re-identification risk and information loss (as shown in Table 1). After calculating the results for the three different de-identification approaches, the Privacy Tailor will choose to perform de-identification on "region" because it has a relatively low re-identification risk and the lowest data distortion level. After finishing this step, the Local re-identification risk will change from 0.73 to the after de-identification risk value 0.23. The Aggregate risk value will union S_1 to S_2. In other words, it rescans the QIDs in the union of S_1 and S_2 to obtain an aggregate risk value of 0.24; Local IL equals 30%, and Aggregate IL equals the sum of IL and that for S_1, which is 0% plus 30%, equals 30%.

(iii) After finishing the assessment of S_2, it will calculate the re-identification risk of the (age × region × sex) combination at S_3, and the result obtained is 0.1, which is lower than the threshold value. After rescanning the union of QIDs in the 3 nodes from S_1 to S_3, the aggregate risk value becomes 0.3 (less than the threshold 0.5). Therefore, the Privacy Tailor will stop de-identification at this level.

This example demonstrates that the Privacy Tailor decides whether to perform de-identification based on the risk level, and then locate the optimal QID information combination from different conditions; de-identification is not performed on all QID information. This multilevel method only needs to deal with local information combinations most of the time and therefore can effectively reduce IL value. In addition, it can also identify the high-risk data in a database and help improve privacy safeguards.

3. Simulation and Results

This section presents a discussion of the experiments performed. The environment developed in C language is used to simulate the workflow of the HT system. We used two datasets in the experiment. The first dataset is sourced from the Microdata (demodata.asl) and Macrodata (demodata.rda) of μ-Argus [16], and is called Dataset 1 (shown with solid lines). The second dataset is sourced from the adult data set of the UCI Machine Learning Repository [17], and is called Dataset 2 (shown using dashed lines). Under the considerations of the re-identification risk threshold between $k = 2$ and $k = 15$, the target attributes are age, address, and income.

Based on assumptions above, the ECG composer outputs an Execution Chain Graph with accessing three QID attributes: age, address, and income. In each node, the Privacy Tailor assesses whether the re-identification risk is higher than the threshold. If the risk is within an acceptable range, the information will be passed to the next node without de-identifying the attribute. In our experiment, the risk values assessed in node one and node two are lower than the threshold, while the node three assessment result is higher than the threshold. Therefore, an appropriate de-identification method combination is required.

Table 1: Different cases in re-indemnification process.

Case	Re-indemnification risk	Information loss
Age	0.55	50%
Region	0.23	30%
Age × Region	0.36	70%

Firstly, the risk of each de-identification combination of the attributes needs to be assessed. There are seven possible de-identification combinations: address, age, income, address × age, age × income, address × income, and address × age × income. When the risk values of all nodes are lower than the threshold, we perform data de-identification with only some of the attributes, which result in low information distortion. The following paragraphs present the results plotted from the experiments. The HT system uses the same de-identification techniques as μ-Argus. With the same re-identification risk threshold (k), we compared the distortion levels between de-identifying with the optimal combination of HT and de-identifying with the entire dataset of μ-Argus. The distortion level is represented by Modification Rate (MR) and Extended Bias In Mean (EBIM).

3.1. Modification Rate. MR represents the distortion level based on the amount of data being modified. The idea here is that when executing a de-identification procedure, a portion of the data is modified, which causes data distortion. Equation (2) is to calculate the ratio between the numbers of modified attributes and the total attribute numbers.

$$\mathrm{MR} = \frac{N_A}{N_T}, \tag{2}$$

where N_A is the number of modified attributes of a dataset, and N_T is the total number of attributes in the dataset.

Figure 4 demonstrates the MR of both the HT system and the μ-Argus system. The x-axis represents the re-identification risk k, and the y-axis represents the MR of the de-identified dataset. As shown in the figure, for Dataset 1, the amount of data that needs to be modified is 65% and 95% for the HT system and μ-Argus system, respectively. According to (2), the distortion level is determined by the amount of data that is modified. Thus, the distortion level of the HT system is 30% lower than that of the μ-Argus system. For Dataset 2, we find that when $k = 2$, the amount of data that needs to be modified is 28% and 70% for the HT system and μ-Argus system, respectively. As the threshold increases, a larger part of dataset needs to be modified, and our system maintains a relatively low-distortion level. Even when $k = 4$, the MR of HT system increases, but remains lower than μ-Argus. Therefore, in terms of MR, the HT system is superior.

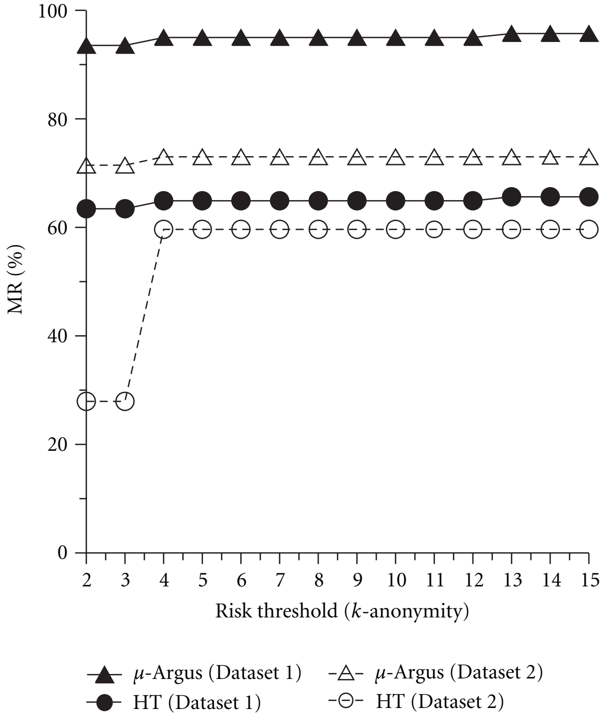


Figure 4: Data distortion analysis on Modification rate.

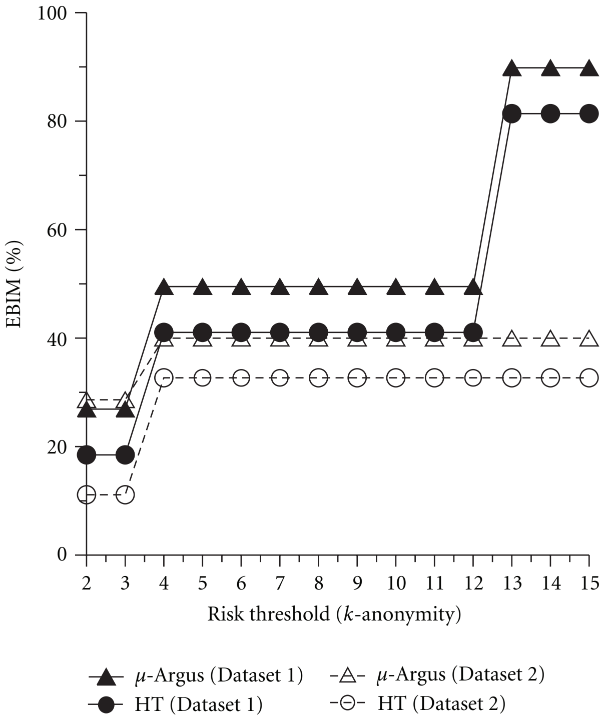


Figure 5: Data distortion on Extended Bias in Mean.

3.2. Extended Bias in Mean. EBIM extends the Bias In Mean (BIM) method, proposed by Li and Sarkar [18], to calculate the difference between the modified dataset and the original dataset. As BIM is only suitable for calculating the difference of single attribute between the modified dataset and the original dataset, the EBIM improved the BIM method to calculate the average of the difference for all attribute fields, before and after modification. To clearly indicate the information loss, we used an extended BIM (EBIM) to accommodate for the generalization strategy. Assuming the interval where the attribute (X) resides is known, the range $R \leq L, X, U >$ where U is the upper bound value; L is the lower bound value; X is the original value. The EBIM formula is given in (3) where j represents the index of the attributes and i represents the index of data entry.

$$\text{EBIM} = \frac{1}{N_T}\sum_{i=1}^{N_T}\sum_{j=1}^{N_A}\left(\frac{U-L}{X_{i,j}}\right), \tag{3}$$

where N_A is the total attribute numbers of a dataset; N_T is the total number of data entries.

As shown in Figure 5, it shows the comparison of the distortion level by EBIM between the HT system and μ-Argus system. The x-axis is the re-identification risk threshold (k). The y-axis represents the EBIM distortion level. Figure 5, presents that the HT system outperforms the μ-Argus system in all scenarios. In Dataset 1, the distortion rate increases as the threshold increases. When $k = 4$, the distortion increases due to the higher level of de-identification required. However, the HT system still manages a lower-distortion level than μ-Argus does. After the previous de-identification is processed, no additional de-identification is required between $k = 4$ and $k = 12$ in Dataset 1 (i.e., remaining the same EBIM results). When $k = 13$ in Dataset 1, both systems should further de-identify data and yielded higher distortion levels. Moreover, in Dataset 2, HT system is able to maintain a lower-distortion level than μ-Argus. Further, no additional de-identification is required beyond $k = 4$ in Dataset 2. Based on both datasets, the HT system produced a comparatively lower-distortion level.

4. Conclusion and Future Work

Safeguarding privacy has received increased attention from the public. Using personal information, we may be able to identify a particular person directly or indirectly. Traditional methods, which perform de-identification on the entire database, can reduce the re-identification risk and protect private information, but they cannot provide authentic information to researchers. Based on experimental results, this paper proposes the HT system, which maintains a low re-identification risk in the required area, but is still able to effectively reduce the level of information loss and satisfy the needs of medical and research groups, and identify the information with high risk. HT system enables administrators to completely customize a privacy-preserved database system for eHealth applications and ensure that all service requests are managed in a consistent and reliable manner. In future work, we will satisfy l-diversity requirement [19] to ensure that sensitive attribute values in each equivalence class are sufficiently diverse in order to make the HT system have more practical privacy protection.

References

[1] J.-H. Kao, C.-Y. Hsu, Y.-P. Sung, and W. P. Liao, "DICOM-based multi-center electronic medical records management system," *International Journal of Bio-Science and Bio-Technology*, vol. 2, no. 2, pp. 11–22, 2010.

[2] S.-H. Lin, Y.-C. G. Lee, and C.-Y. Hsu, "Data warehouse approach to build a decision-support platform for orthopedics based on clinical and academic requirements," *International Journal of Bio-Science and Bio-Technology*, vol. 2, no. 1, pp. 1–12, 2010.

[3] J. Pedraza, M. A. Patricio, A. de Asís, and J. M. Molina, "Privacy and legal requirements for developing biometric identification software in context-based applications," *International Journal of Bio-Science and Bio-Technology*, vol. 2, no. 1, pp. 13–24, 2010.

[4] Health System Use Technical Advisory Committee—Data De-Identification Working Group, "'Best Practice' Guidelines for Managing the Disclosure of De-Identified Health Information," Ottawa, Canada, Canadian Institute for Health Information, 2010.

[5] K. El Emam, "Risk-based de-identification of health data," *IEEE Security and Privacy*, vol. 8, no. 3, pp. 64–67, 2010.

[6] K. El Emam, "Heuristics for de-identifying health data," *IEEE Security and Privacy*, vol. 6, no. 4, pp. 58–61, 2008.

[7] A. Appari and M. E. Johnson, "Information security and privacy in healthcare: current state of research," *International Journal of Internet and Enterprise Management*, vol. 6, no. 4, 2010.

[8] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[9] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*, vol. 10, no. 5, pp. 571–588, 2002.

[10] K. El Emam and F. K. Dankar, "Protecting Privacy Using k-Anonymity," *Journal of the American Medical Informatics Association*, vol. 15, no. 5, pp. 627–637, 2008.

[11] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," in *Proceedings of the IEEE Symposium on Research in Security and Privacy*, Oakland, Calif, USA, May 1998.

[12] R. Fraser and D. Willison, "Tools for De-Identification of Personal Health Information," Pan Canadian Health Information Privacy (HIP) Group, 2009.

[13] K. El Emam, F. K. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk, "Evaluating the risk of re-identification of patients from hospital prescription records," *Canadian Journal of Hospital Pharmacy*, vol. 62, no. 4, pp. 307–319, 2009.

[14] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *Proceedings of the 21st International Conference on Data Engineering (ICDE '05)*, pp. 205–216, Tokyo, Japan, April 2005.

[15] F. K. Dankar and K. El Emam, "A method for evaluating marketer re-identification risk," in *Proceedings of the EDBT/ICDT Workshops*, Lausanne, Switzerland, March 2010.

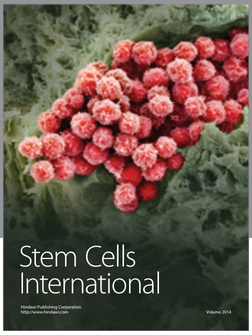
[16] Voorburg Group, "μ-Argus version 4.2 Software and User's Manual," Netherlands Statistical Office, 2008.

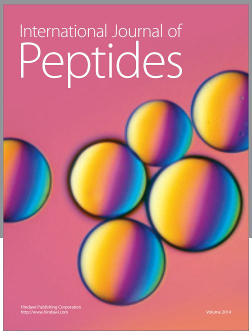
[17] A. Frank and A. Asuncion, "UCI Machine Learning Repository," University of California, School of Information and Computer Science, 2010, http://archive.ics.uci.edu/ml.

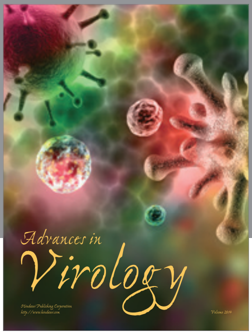
[18] X. B. Li and S. Sarkar, "A tree-based data perturbation approach for privacy-preserving data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 9, pp. 1278–1283, 2006.

[19] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, Article ID 1217302, 2007.

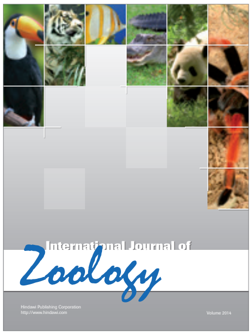
BioMed
Research International

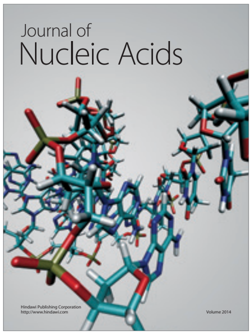

Stem Cells
International

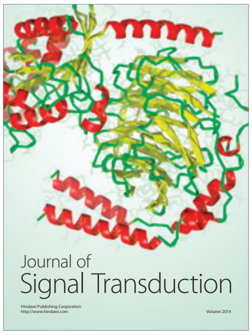

International Journal of
Peptides


Advances in
Virology

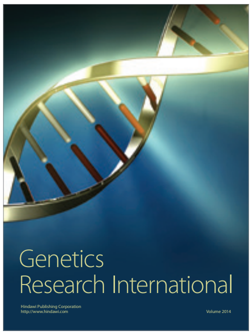
International Journal of
Genomics

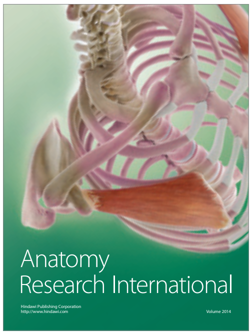

International Journal of
Zoology


Journal of
Nucleic Acids


Journal of
Signal Transduction

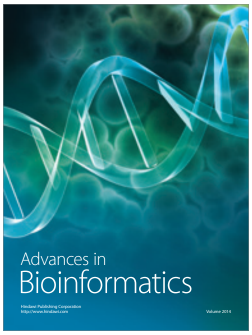
The Scientific
World Journal

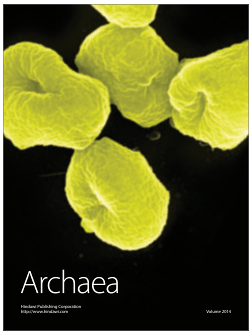

Genetics
Research International

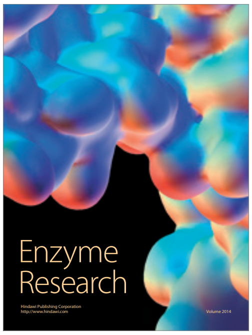

Anatomy
Research International

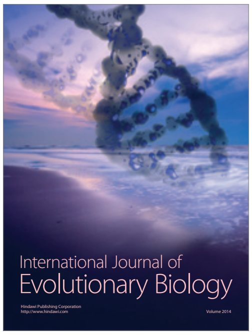
International Journal of
Microbiology

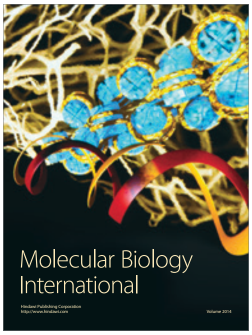
Biochemistry
Research International


Advances in
Bioinformatics


Archaea


Enzyme
Research


International Journal of
Evolutionary Biology


Molecular Biology
International

Journal of
Marine Biology