Original Paper

Outsourcing Medical Data Analyses: Can Technology Overcome Legal, Privacy, and Confidentiality Issues?

Bostjan Brumen[1*], BCompSci, PhD; Marjan Heričko[1*], BCompSci, MSc, PhD; Andrej Sevčnikar[1*], BCompSci; Jernej Završnik[2*], MD, MSc; Marko Hölbl[1*], BCompSci, PhD

[1]Institute of Informatics, Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia

[2]Health Care Center Maribor, Maribor, Slovenia

[*]all authors contributed equally

Corresponding Author:
Marko Hölbl, BCompSci, PhD
Institute of Informatics
Faculty of Electrical Engineering and Computer Science
University of Maribor
FERI G2
Smetanova 17
Maribor, 2000
Slovenia
Phone: 386 2 2207292
Fax: 386 2 2207272
Email: marko.holbl@uni-mb.si

Abstract

Background: Medical data are gold mines for deriving the knowledge that could change the course of a single patient’s life or even the health of the entire population. A data analyst needs to have full access to relevant data, but full access may be denied by privacy and confidentiality of medical data legal regulations, especially when the data analyst is not affiliated with the data owner.

Objective: Our first objective was to analyze the privacy and confidentiality issues and the associated regulations pertaining to medical data, and to identify technologies to properly address these issues. Our second objective was to develop a procedure to protect medical data in such a way that the outsourced analyst would be capable of doing analyses on protected data and the results would be comparable, if not the same, as if they had been done on the original data. Specifically, our hypothesis was there would not be a difference between the outsourced decision trees built on encrypted data and the ones built on original data.

Methods: Using formal definitions, we developed an algorithm to protect medical data for outsourced analyses. The algorithm was applied to publicly available datasets (N=30) from the medical and life sciences fields. The analyses were performed on the original and the protected datasets and the results of the analyses were compared. Bootstrapped paired *t* tests for 2 dependent samples were used to test whether the mean differences in size, number of leaves, and the accuracy of the original and the encrypted decision trees were significantly different.

Results: The decision trees built on encrypted data were virtually the same as those built on original data. Out of 30 datasets, 100% of the trees had identical accuracy. The size of a tree and the number of leaves was different only once (1/30, 3%, *P*=.19).

Conclusions: The proposed algorithm encrypts a file with plain text medical data into an encrypted file with the data protected in such a way that external data analyses are still possible. The results show that the results of analyses on original and on protected data are identical or comparably similar. The approach addresses the privacy and confidentiality issues that arise with medical data and is adherent to strict legal rules in the United States and Europe regarding the processing of the medical data.

(J Med Internet Res 2013;15(12):e283) doi: 10.2196/jmir.2471

KEYWORDS

confidentiality; patient data privacy; data protection; medical decision making; computer-assisted; data analysis

Introduction

Background

Medical data are gold mines for deriving knowledge. Hiding within those mounds of data is knowledge that could change the life of a single patient, or sometimes change the health of an entire population [1]. Medical doctors—field experts—use the data collected from various sources on a daily basis for treating patients. Many data from examinations and laboratory tests require further analyses, which can be very time consuming and require experts to conduct them. In fact, the amount of data produced by medical electronic equipment is enormous and continues to grow at a very fast rate—the amount of data doubles in approximately 15 months [1-3]. The large volumes of data make human-driven analyses impossible. Machine support and intelligent data analyses are definitely required. Medical experts and other employees of a health care service provider generally do not possess in-house expertise for doing automatic data analyses. There is also a distinction between deriving knowledge from Web pages, blogs, social media systems, etc, and from the closed systems typically present in medical environments. The former is mostly used for branding purposes (advertising, marketing, and content delivery) [4] and the latter to support doctor's decision making. Sometimes information technology (IT)-related resources in a health care system, including hardware and software, may not be adequate or even available for the data analyses aimed at knowledge discovery. The obvious choice is to have a third party conduct the analyses.

Here, privacy and confidentiality issues arise together with legal obligations. Respect for privacy has been a part of the medical profession since ancient times. "Whatever I see or hear in the lives of my patients, whether in connection with my professional practice or not, which ought not to be spoken of outside, I will keep secret, as considering all such things to be private..." is the text from an oath attributed to Hippocrates referring to confidentiality [5]. Privacy and confidentiality are very important contemporary issues, especially in the Western world, and are not limited to the medical field.

Privacy has (re)emerged as an important issue since the emergence of social media, as noted by Mark Zukerberg, the founder of the most-used social network, Facebook [6], and Facebook's chief operation officer, Sheryl Sandberg. They observed that privacy controls were centered at Facebook's core at all times [7,8]. Indeed, privacy needs to be considered seriously from a technological point of view when designing applications and solutions. In Canada, the Ontario Privacy Commissioner, Ann Cavoukian, has developed a Privacy by Design (PbD) framework [9-11] which emphasizes the need to adopt a proactive rather than a reactive compliance approach to the protection of privacy.

The laws of most developed countries impose obligations to respect informational privacy (eg, confidentiality, anonymity, secrecy, and data security), physical privacy (eg, modesty and bodily integrity), associational privacy (eg, intimate sharing of death, illness, and recovery), proprietary privacy (eg, self-ownership and control over personal identifiers, genetic data, and body tissues), and decisional privacy (eg, autonomy and choice in medical decision making) [12,13]. In this paper, however, we address the first type of privacy: informational privacy.

Informational privacy is usually violated by a data breach, which can result from theft, intentional or accidental unauthorized access to data, acts of revenge by unsatisfied employees, or by the accidental loss of media or devices that bear data.

Despite the regulations in place, the stories of privacy and confidentiality breaches are still frequent. Major hospitals and health-related institutions, most notably in the United States but also elsewhere in the world, have experienced highly publicized data breaches—more than 770 breaches have occurred since 2005 in the United States alone [14]. Anciaux et al [15] observed that traditional electronic health records (EHR) have no security guarantee outside the health care service domain and pervasive health, a new concept based on latest developments, requires implementable principles for privacy and trustworthiness [16]. Van der Linden et al [17] noticed that before the virtual lifelong patient record can become reality, more clarity has to be provided on the legal and computational frameworks that protect confidentiality.

Can technology help and how can it help? First, let us take a closer look at definitions of privacy and confidentiality and how they are reflected in laws and rules.

Privacy and Confidentiality

Daniel Solove [18] has stated: "Privacy is a concept in disarray. Nobody can articulate what it means." But one must note that privacy and confidentiality do share at least some common grounds among the philosophers and jurists, and many technologies exist that address privacy and confidentiality.

In Ancient Greek civilization, there existed 2 interdependent and sometimes conflicting areas: the public area of politics and political activity, the *polis*, and the private area of the family, the *oikos* [19,20]. These areas were reflected in classic dramas (eg, in Sophocles' *Antigone* and *Oedipus Rex*), and the new order of the polis, despite its weaknesses, reigned supreme at the end of the dramas [21].

More systematic discussion of the concept of privacy began with an article by Samuel Warren and Louis Brandeis titled "The Right to Privacy" [22]. Citing "political, social, and economic changes" and a recognition of "the right to be let alone," they argued that existing laws afforded a way to protect the privacy of the individual, and they sought to explain the nature and extent of that protection. Focusing in large part on the press and publicity allowed by recent inventions, such as photography and newspapers, but referring to violations in other contexts as well, they emphasized the invasion of privacy brought about by public dissemination of details relating to a person's private life. Warren and Brandeis felt a variety of existing cases could be protected under a more general right to privacy which would protect the extent to which one's thoughts, sentiments, and emotions could be shared with others. They were not attempting to protect the items produced or intellectual property, but rather the peace of mind attained with such protection; they said the right to privacy was based on a principle of "inviolate personality" which was part of a general right of immunity of the person: “the right to one’s personality” [22]. Thus, Warren and Brandeis laid the legal foundation for a concept of privacy that has come to be known as control over information about oneself [23].

In an attempt to systematize and more clearly describe and define the new right of privacy upheld in tort law, William Prosser [24] wrote in 1960 about 4 different interests in privacy, or privacy rights:

1. Intrusion upon a person’s seclusion or solitude, or into his private affairs;
2. Public disclosure of embarrassing private facts about an individual;
3. Publicity placing one in a false light in the public eye; and
4. Appropriation of one’s likeness for the advantage of another [23].

Prosser noted that the intrusion in the first privacy right had expanded beyond physical intrusion, and pointed out that Warren and Brandeis had been concerned primarily with the second privacy right. Nevertheless, Prosser felt that both real abuses and public demand had led to general acceptance of these 4 types of privacy invasions. Thomas Nagel, one of the America’s top contemporary philosophers, gives a more contemporary (philosophical) discussion of privacy, concealment, publicity, and exposure [25].

More recently, Adam Moore [26], building on the views of Ruth Gavison [27], Anita Allen [28], Sissela Bok [29], and others offered a control-over-access account of privacy. According to Moore, privacy is a cultural- and species-relative right to a level of control over access to bodies or places of information. While defending the view that privacy is relative to species and culture, Moore argues that privacy is objectively valuable: human beings that do not obtain a certain level of control over access will suffer in various ways. Moore claims that privacy, like education, health, and maintaining social relationships, is an essential part of human flourishing or well-being [23].

In a medical context, as viewed by Allen [13], the privacy at issue is very often confidentiality [30], specifically the confidentiality of patient-provider encounters (including the fact that an encounter has taken place), along with the secrecy and security of information memorialized in physical, electronic, and graphic records created as a consequence of these encounters [30]. Confidentiality is defined as restricting information to persons belonging to a set of specifically authorized recipients [13,28,31,32]. Confidentiality can be achieved through either professional silence, leaning on the moral aspect, or through secure data management [33], leaning on technologies and techniques.

The moral significance attached to medical privacy is reflected in data protection and security laws adopted by local and national authorities around the world. The point of these laws is to regulate the collection, quality, storing, sharing, and retention of health data, including the EHR [13].

Medical Data Legal Regulations in the United States and Europe

United States

In the United States, several prominent cases in the 1990s aroused public and legal interest in privacy and confidentiality of medical data. There was no federal law regulating privacy and confidentiality before 1996. One of the key turning points was a breach of Nydia Velasquez’s medical records during her campaign for a House seat. At hearings before the US Senate Subcommittee on Technology and the Law of the Committee on the Judiciary on January 27, 1994, she said:

> *...I woke up one morning with a phone call from my friend Pete Hamill, a columnist at the New York Post. He told me that the night before, the Post had received an anonymous fax of my records from St. Claire Hospital. The records showed that I had been admitted to the hospital a year ago seeking medical assistance for a suicide attempt. He told me that other newspapers across the city had received the same information, and the New York Post was going to run a front page story the next day. For the press, it was a big story. For me, it was a humiliating experience over which I had no control...When I found out that this information was being published in the newspaper and that I had no power to stop it, I felt violated. I trusted the system and it failed me. What is most distressing is that once medical records leave the doctor’s office, there are no Federal protections to guard against the release of that information. In some States, it is easier to access a person’s medical history than it is to obtain the records of a person’s video rentals... [34]*

Many similar stories have urged US legislators to adopt federal regulations implemented under the Health Insurance Portability and Accountability Act (HIPAA) of 1996 [35]. Before the HIPAA, no generally accepted set of security standards or general requirements for protecting health information existed in the health care industry. Under HIPPA, the US Department of Health and Human Services (HHS) has adopted 5 administrative rules, among them the HIPAA Privacy Rule [36] and the HIPAA Security Rule [37], the latter complementing the former. The Privacy Rule deals with all protected health information (PHI) regardless of the form (ie, including paper and electronic formats), and the Security Rule deals specifically with electronic PHI (ePHI).

The HIPAA Privacy Rule, or the *Standards for Privacy of Individually Identifiable Health Information*, is a set of federal standards to establish protection of certain health information. The *Security Standards for the Protection of Electronic Protected Health Information* (the Security Rule) established a national set of security standards for protecting certain health information that is held or transferred in electronic form. The Security Rule operationalizes the protections contained in the Privacy Rule by addressing the technical and nontechnical safeguards that organizations, called *covered entities*, must put in place to secure individuals’ ePHI. [38]. The Security Rule specifies administrative, technical, and physical measures that must be adopted by covered entities to adequately protect the privacy and confidentiality of ePHI.

Additionally, the HHS issued a set of rules [39] requiring the covered entities to notify individuals when their health information is breached. Furthermore, the covered entity must inform the HHS Secretary and the media when a breach involves more than 500 persons; thus, implementing provisions of the Health Information Technology for Economic and Clinical Health (HITECH) Act. The rules also apply to the business associates of the covered entities to notify the covered entity of events that affect privacy and confidentiality of ePHI at or by the business associate.

In the Breach Notification Rule [39], the HHS has specified the encryption and destruction as the technologies and methodologies that render PHI unusable, unreadable, or indecipherable to unauthorized individuals. Entities subject to the HHS and Federal Trade Commission regulations that secure health information as specified by the guidance through encryption or destruction are relieved from having to notify in the event of a breach of such information [39,40].

Europe

In 1995, the European Parliament passed Directive 95/46/EC on the protection of individuals in regard to the processing of personal data and the free movement of such data [41]. Member States in the European Union can, within the limits of the provisions of the Directive, determine more precisely the conditions under which the processing of personal data is lawful. Based on the Directive, the European Parliament and the Council on December 18, 2000, adopted the Regulation (EC) No 45/2001 on the protection of individuals in regard to the processing of personal data by the Community institutions and bodies and the free movement of such data [42].

Interestingly, Article 8 of the Directive 95/46/EC explicitly prohibits the processing of special categories of data, including the processing of data concerning health. However, the prohibition does not apply where processing of the data is required for the purposes of preventive medicine, medical diagnosis, the provision of care or treatment, or the management of health care services, and where those data are processed by a health professional subject under national law or rules established by national competent bodies to the obligation of professional secrecy or by another person also subject to an equivalent obligation of secrecy [41].

Furthermore, Article 17 of the directive prescribes security of processing. The controller of data must implement appropriate technical and organizational measures to protect personal data against accidental or unlawful destruction or accidental loss, alteration, or unauthorized disclosure or access, in particular where the processing involves the transmission of data over a network, and against all other unlawful forms of processing. Having regard to the state of the art and the cost of their implementation, such measures shall ensure a level of security appropriate to the risks represented by the processing and the nature of the data to be protected. The controller must, when processing is carried out on his behalf, choose a processor providing sufficient guarantees in respect of the technical security measures and organizational measures governing the processing to be carried out, and must ensure compliance with those measures. Processing by way of a processor must be governed by a contract or legal act binding the processor to the controller, stipulating that (1) the processor shall act only on instructions from the controller, and (2) the obligations regarding the appropriate technical and organizational measures to protect personal data, as defined by the law of the Member State in which the processor is established, shall also be incumbent on the processor. The contract or the legal act between the controller and the processor relating to data protection and the appropriate technical and organizational measures to protect personal data must be in written form [41].

When personal data are processed by automated means, measures shall be taken as appropriate in view of the risks. The measures should ensure that during communication of personal data and during transport of storage media, the data cannot be read, copied, or erased without authorization [42].

Directive 95/46/EC has been unchanged in principle since 1995. At the beginning of 2012, the European Commission proposed a comprehensive reform of the 1995 data protection rules to strengthen online privacy rights and boost Europe’s digital economy. Technological progress and globalization have profoundly changed the way the data are collected, accessed, and used. In addition, the 27 EU Member States have implemented the 1995 rules differently, resulting in divergences in enforcement. The proposed law would restrict the way Internet companies can gather, use, and retain the volumes of personal data that their users post online [43]. Among other measures, the use of encryption standards may be required in certain situations (Article 27), and a 24-hour notification rule is proposed: in a case of a personal data breach, the controller must notify, without undue delay and, when feasible, not later than 24 hours after having become aware of it, the personal data breach to the supervisory authority (Article 28). The European regulation, once passed, could serve as a template for other countries as they draft or revise their data protection policies [44], and it is threatening the current business practices of the Internet giants, such as Facebook [45].

Technological Similarities in Protecting Medical Data in the United States and Europe

The main difference between the American and European legislation pertaining to medical data is in the level of detail of how the data should be protected. The HIPAA and the accompanying rules, especially the Privacy and Security Rules, give great detail in how to protect data. In Europe, the detail of the protection is left to the EU Member States who must apply national provisions pursuant to Directive 95/46/EC within 3 years from the adoption of the directive.

However, there is one common point: both systems suggest the use of encryption to protect sensitive data. Although the HIPAA Security Rule does not dictate the use of encryption, it becomes an evident choice when considering the HITECH Breach Notification Rule. Entities covered by the rule are relieved from having to notify the media and others in the event of a breach of encrypted information. The EU Regulation (EC) No 45/2001, based on the Directive 95/46/EC, suggests the use of encryption when processing data for historical, statistical, or scientific purposes (Article 4) [42]. On the other hand, local laws of EU Member States usually do not dictate the use of encryption, as in case of the Data Protection Act of 1998 in the United Kingdom [46]. The same, for example, is true for the German Federal Act on Protection of Data [47]. It seems that recommendations to use encryption are lowered to the level of various guidance and recommendations [48,49].

Regardless of the legal system and local rules, the use of encryption seems an obvious choice for protecting medical data.

Technology to Increase Confidentiality and Privacy With Outsourced Data Analyses

The fact that the outsourced data analyses poses a potential security threat to data has been well known for decades [50]. To protect sensitive data, several techniques have been developed.

Firstly, the techniques developed for the protection of statistical databases can be used. The goal of these techniques is to disclose the statistical data (eg, sums, counts, averages, minimums, maximums) without exposing sensitive individual records [51]. In these cases, the sensitive individual data values are either generalized or not disclosed. In the data analysis world, we cannot have data that have been generalized or are not available at all.

A typical result of an intelligent data analysis is a set of decision rules. A decision rule is a function which maps an observation to an appropriate action. Such rules are typically found in a medical diagnosis process in which several measurements are observed and an action is taken (eg, a drug is prescribed). For example, a computer-generated decision rule on generalized and not disclosed data would read: if a patient’s 2-hour postload plasma glucose level is ≤199 mg/dL and the patient’s body mass index (BMI) and age are unknown, then diagnosis of diabetes mellitus is negative. However, the American Diabetes Association recommends a postload glucose level ≤155 mg/dL with a 75 g glucose load [52]. High values may indicate diabetes and the doctors will not use just a single test result (measurement) to diagnose diabetes mellitus. If a doctor receives nondisclosed data (or a rule created on nondisclosed data) from the computer-assisted decision system, she has no use of it because additional data are needed for the final decision.

Secondly, one can modify the real value of an attribute using a value-class membership technique or value distortion [53] and try to reconstruct the original distribution as close as possible [54]. In the first case, the values are partitioned into a set of disjointed, mutually exclusive classes; for example, the numeric value of 2-hour postload glucose can be divided into 3 separate disjointed classes (c), 0-139 mg/dL, 140-199 mg/dL, and 200-299 mg/L, written as c_1 = (0..139), c_2 = (140..199), and c_3 = (200..999), respectively. The selection of classes needs to be done carefully based on the domain knowledge; otherwise, the approach is useless. In the second case, the values are slightly changed, namely a random value drawn from some distribution is added to the original value. This approach can be used for numeric attributes (only) and for constructing a classifier [53]. Previous research focused on cases in which the data were distorted, expecting that the data were (deliberately) changed at the entry point into a system. In many cases, the models built were very sensitive to distorted values. For example, a computer-generated decision rule on distorted data may read: if the 2-hour postload plasma glucose level is ≤150 mg/dL and the BMI is >35 kg/m^2 and age is ≤35 years, then diabetes mellitus diagnosis is negative, instead of the original if the 2-hour postload plasma glucose level is ≤127 mg/dL and BMI >26.4 kg/m^2 and age is ≤28 years, then diabetes mellitus diagnosis is negative. An action based on a wrong decision rule can have serious consequences, especially in cases in which the values are very sensitive to small changes. However, the mentioned works are orthogonal to the work presented in this paper and can be used complementarily, if needed.

Thirdly, the encryption techniques can be implemented so that the data are encrypted on-site before they are sent for analysis. An analyst decrypts the data based on a password that was previously agreed upon and works with the original values. Typically, a data owner stores the data in an Excel or Word file and protects it by using the internal protection methods; alternatively, the data are stored in another format and compressed using WinZip tools, again protecting it with an internal protection method. The files are then transported to the outside world. Such a procedure has many drawbacks. Firstly, the data are not protected once the outsourced external analyst receives the files and deactivates the files’ internal protection to access the data. The data are vulnerable to any and all attacks possible once residing on the analyst’s computer. Secondly, the password with which the files are protected can easily be broken. A recent study showed that 93% of test files containing sensitive medical data could be recovered within a 24-hour period by using commercially available tools [55]. Interestingly, nothing has changed in the terms of using strong passwords for decades [56,57]. It can be concluded that passwords will continue to be the weakness of computing security.

The Contribution

The aim of the present work is to develop a procedure to protect medical data in such a way that the outsourced analyst is capable of doing analyses on protected data and the results will be comparable, if not the same, as if they had been done on the original data by following the PbD principle. We tested this hypothesis by determining whether there were differences between outsourced decision trees built on encrypted data and the ones built on original data.

Methods

Formal Setting for Encrypting Data for Outsourced Analyses

In our proposed method, we avoided the weaknesses of the previously mentioned approaches. The data values were encrypted in such a way that outsourced data analyses were still possible, but the data remain encrypted and protected. This can be done by using a strong encryption algorithm, such as those approved by National Institute of Standards and Technology (NIST): Triple Data Encryption Algorithm (TDEA) [58], Advanced Encryption Standard (AES) [59], or Skipjack [60], so that the security should rely only on secrecy of the keys [61].

The formalization of the approach is presented in Multimedia Appendix 1 and is based on the flat file format (in principle, a textual file with data items separated by a comma), which is the usual format for data analytic tools [62].

The Algorithm for Protection of Data for Decision-Making Analyses

We designed an algorithm that encrypts a flat file with plain text data into an encrypted flat file in such a way that external data analyses are still possible. The algorithm is presented in Multimedia Appendix 2.

For clarity of the proposed approach, let us take a closer look at an experiment with real-world examples from the medical and life sciences fields.

Data Collection

For the purpose of demonstrating the usability of the proposed approach, we used all publicly available datasets from the University of California at Irvine (UCI) Machine Learning Repository [63], with the following restriction: the problem task was classification, data type was multivariate, from the life sciences area, and the data were in matrix (table) format. The UCI Machine Learning Repository lists 41 such datasets [64]. We further removed the following 11 datasets: Arcene, Dorothea, and p53 Mutants (the number of attributes >1000, the primary task is feature selection, not classification), both of the Kyoto Encyclopedia of Genes and Genomes (KEGG) datasets and the PubChem Bioassay Database (textual data), Parkinson’s (time series data), and the Thyroid Disease family of datasets (the task is from domain theory). Next, we used only original or larger datasets in which several sub-datasets were available (removed Breast Cancer Wisconsin Diagnostic and Prognostic, Soybean-small, SPECT Heart). We ended up experimenting with 30 datasets.

Most of the datasets in Attribute-Relation File Format (ARFF) were taken from the Software Environment for the Advancement of Scholarly Research (SEASR) repository [65], the rest were converted to ARFF from the UCI repository files by the authors. The original ARFF files are included in Multimedia Appendix 3.

Data Processing

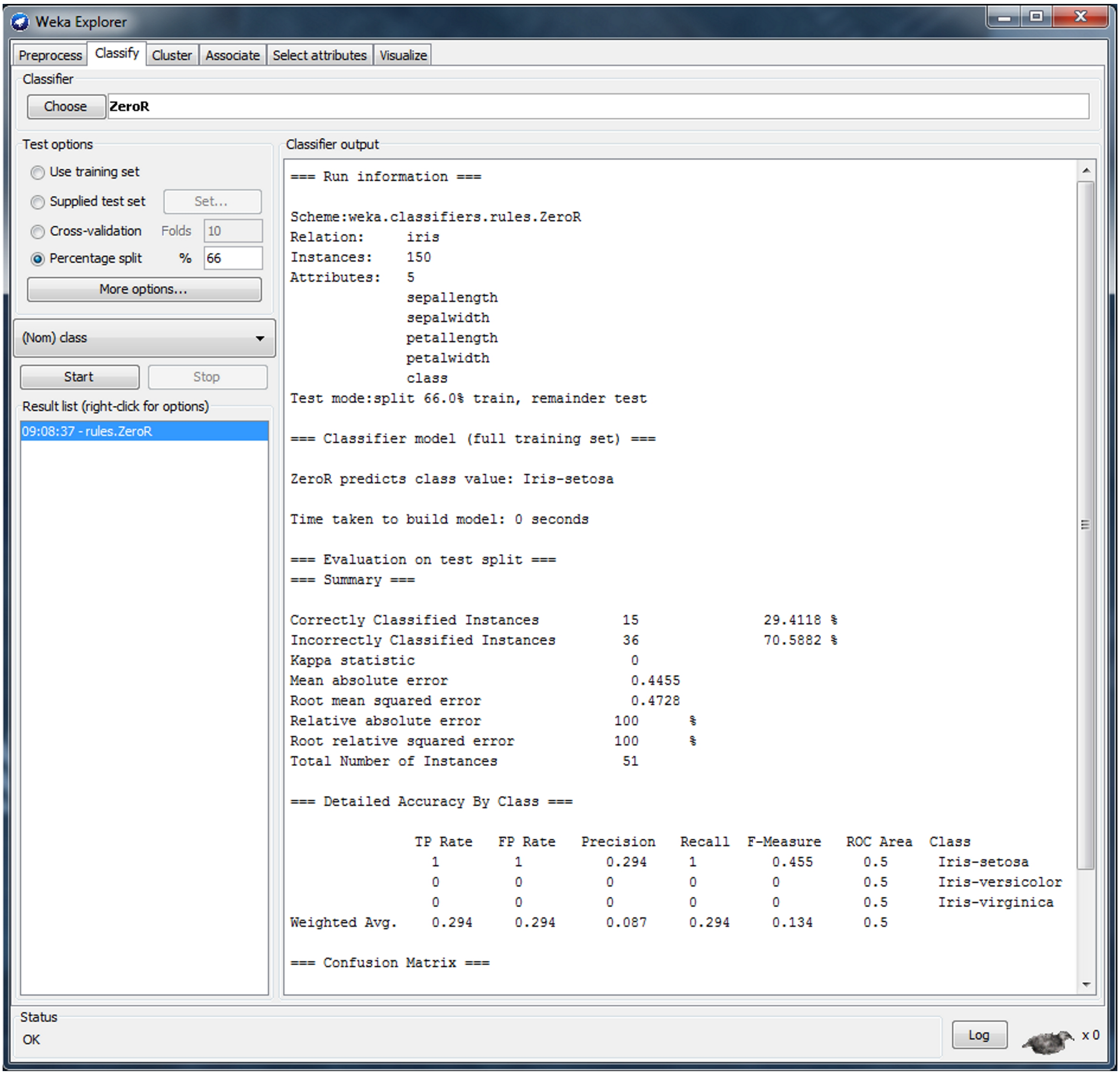
For the analytics tool in this experiment, we chose the J48 decision tree builder with standard built-in settings and initial values, which is freely available from the Waikato Environment for Knowledge Analysis (Weka) project toolkit [62] version 3.6.8. J48 is java-based decision tree builder based on a Quinlan’s C4.5 tree induction [66].

First, each original dataset was used to build a decision tree using the J48 decision tree (see Figure 1). We used 66% of all dataset items for training and the remaining data were used for testing the model; therefore, we ignored any separate training or test set, or any associated cost model.

For each decision tree model, we measured the number of leaves, size of the tree, and the percentage of correctly classified instances (see Multimedia Appendix 4). The number of leaves defines the total number of decision rules included in a tree. The size of a tree gives the number of nodes (measurements) in a tree: the higher the number of nodes, the more complicated the rules are. The percentage of correctly classified instances (ie, accuracy) measures how many mistakes the computer-generated decision trees make when they are tested on real-world data. Measuring only the accuracy is not enough because many different trees based on different data can have identical accuracy.

Secondly, each data file was protected with the proposed algorithm (see Multimedia Appendix 5). We implemented a prototype with limited features in JavaScript language (see Multimedia Appendix 6). The advantage of using JavaScript is that the data are not sent to a server residing elsewhere, but are processed in a browser locally. We used the AES algorithm with 256-bit key on all string-, categorical-, or nominal-type attribute values. For numeric values, we simply multiplied the original values by 2 and added 1, thus hiding the original values. In real life situations, any numeric transformation preserving the desired statistical properties of data can be used [51]. Then, decision trees were built for each protected dataset with the same settings as the original datasets. Finally, the number of leaves, the size of the tree, and the percentage of correctly classified instances were measured again (see Multimedia Appendix 7).

Figure 1. Weka Explorer user interface showing settings used.



Classifier: ZeroR

Test options

○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

(Nom) class

Result list (right-click for options)

09:08:37 - rules.ZeroR

Classifier output

```
=== Run information ===

Scheme:weka.classifiers.rules.ZeroR
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

ZeroR predicts class value: Iris-setosa

Time taken to build model: 0 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          15               29.4118 %
Incorrectly Classified Instances        36               70.5882 %
Kappa statistic                          0
Mean absolute error                      0.4455
Root mean squared error                  0.4728
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1         0.294     1         0.455      0.5      Iris-setosa
                 0         0         0         0         0          0.5      Iris-versicolor
                 0         0         0         0         0          0.5      Iris-virginica
Weighted Avg.    0.294     0.294     0.087     0.294     0.134      0.5

=== Confusion Matrix ===
```

Status

OK

Hypotheses

For the approach to be useful there should be no statistically significant difference between the original and encrypted trees in terms of tree size, number of leaves in a tree, and the accuracy of the tree. Our hypotheses were (1) the mean tree size after encryption would be the same as before encryption, (2) the mean number of leaves after encryption would be the same as before encryption, and (3) the mean accuracy after encryption would be the same before encryption.

Statistical Analysis

The same subject (a decision tree built with a specific dataset) was observed under 2 different conditions. The “before” samples were made of decision trees built on original data, and the “after” samples were made of decision trees built on encrypted data. Bootstrapped paired *t* tests for 2 dependent samples were used to identify whether significant differences occurred because of encryption of data on 3 independent variables: tree size, number of leaves, and accuracy. We considered differences to be significant at the <.05 level. SPSS version 21 (IBM Corp, Armonk, NY, USA) was used for analysis.

Results

Differences Between Decision Trees Based on Original and Protected Data

First, we tested if the decision trees built on original and protected data were different. Table 1 lists the size of a tree, the number of leaves in a tree, and the percentage of correctly classified items for each dataset when the tree was built on original data and when it was built on the protected data.

Table 1. Results of analyses on original and encrypted data files for tree size, number of leaves, and accuracy.

Database name	Original dataset			Encrypted dataset		
	Tree size[a], n	Leaves[b], n	Accuracy[c], %	Tree size[a], n	Leaves[b], n	Accuracy[d], %
Abalone	2312	1183	21.97	2312	1183	21.97
Acute inflammations	5	3	100.00	5	3	100.00
Arrhythmia	99	50	71.43	99	50	71.43
Audiology (standardized)	54	32	83.12	54	32	83.12
Breast cancer	6	4	68.04	6	4	68.04
Breast cancer Wisconsin (original)	27	14	95.38	27	14	95.38
Breast tissue	29	15	47.22	29	15	47.22
Cardiotocography	19	14	98.34	33	25	98.34
Contraceptive method choice	263	157	55.29	263	157	55.29
Covertype	29,793	14,897	93.59	29,793	14,897	93.59
Dermatology	41	31	92.74	41	31	92.74
Echocardiogram	9	5	70.37	9	5	70.37
Ecoli	43	22	78.95	43	22	78.95
Haberman’s survival	5	3	75.96	5	3	75.96
Hepatitis	21	11	79.25	21	11	79.25
Horse colic	29	18	68.55	29	18	68.55
Iris	9	5	96.08	9	5	96.08
Lung cancer	16	10	63.64	16	10	63.64
Lymphography	34	21	78.00	34	21	78.00
Mammographic mass	15	12	82.26	15	12	82.26
Mushroom	30	25	100.00	30	25	100.00
Pima Indians diabetes	39	20	76.25	39	20	76.25
Post-operative patient	1	1	70.97	1	1	70.97
Primary tumor	88	47	39.13	88	47	39.13
Seeds	15	8	97.18	15	8	97.18
Soybean (large)	93	61	90.52	93	61	90.52
Spectf heart	17	9	66.67	17	9	66.67
Statlog (heart)	45	27	76.09	45	27	76.09
Yeast	369	185	58.81	369	185	58.81
Zoo	17	9	94.12	17	9	94.12

[a] Number of nodes (measurements) in a tree.

[b] Number of decision rules in a tree.

[c] Percentage of correctly classified original items with respect to all items (ie, the number of times the tree's rules lead to the right decision).

[d] Percentage of correctly classified encrypted items with respect to all items (ie, the number of times the tree's rules lead to the right decision).

The analysis of the results showed that all but 1 of the encrypted decision trees were identical to the original ones on all 3 attributes: tree size, number of leaves, and accuracy. The only difference was with the tree built on the Cardiotocography dataset, in which the size of the tree and the number of leaves were different (tree size: 19 vs 33; leaves: 14 vs 25 for original and encrypted datasets, respectively). The difference is due to internals of the algorithm building a decision tree: the algorithm decides how to build the decision tree based on the measurement values; when they are the same, the decision how to build is based on the measurement names, which are not preserved with encryption. Because the values are the same, the induced decision trees are different only in the structure and not in the accuracy or the meaning of the rules.

We tested our hypotheses by using bootstrapped paired samples *t* tests, (see Multimedia Appendix 8). The paired samples results are shown in Table 2.

The unusually high standard deviation indicates the presence of outliers in the data. The outliers are in the Abalone and Covtype data. Outliers tend to increase the estimate of sample variance; thus, decreasing the calculated *t* statistic and lowering the chance of rejecting the null hypothesis. Therefore, we used bootstrapping for the paired samples test, which makes no assumption about underlying population distributions [67]. The results of the bootstrapped paired samples tests are presented in Table 3.

With a significance of *P*=.19, we cannot reject the hypotheses that the mean difference in tree size and in number of leaves would be the same as before encryption. The before and after samples are the same, so we retain the hypothesis that the mean accuracy after encryption would be the same as before encryption.

Table 2. Paired samples statistics.

Pairs	Mean	SD	SEM
Pair 1, n=30			
Original size	1118.1	5432.1	991.8
Encrypted size	1118.6	5432.0	991.8
Pair 2, n=30			
Original leaves	563.5	2715.7	495.8
Encrypted leaves	563.7	2715.7	495.8
Pair 3, n=30			
Original accuracy	0.763[a]	0.189	0.034
Encrypted accuracy	0.763[a]	0.189	0.034

[a] The correlation and *t* test cannot be computed because the standard error of the difference is zero.

Table 3. Bootstrapped paired samples test results.

Pairs	Mean	Bias	SEM	95% CI	*P*
Pair 1: Original size–encrypted size	–0.5	–0.3	0.4	–2.1, –0.5	.19
Pair 2: Original leaves–encrypted leaves	–0.2	–0.1	0.2	–0.9, –0.2	.19

Usability of Encrypted Decision Trees

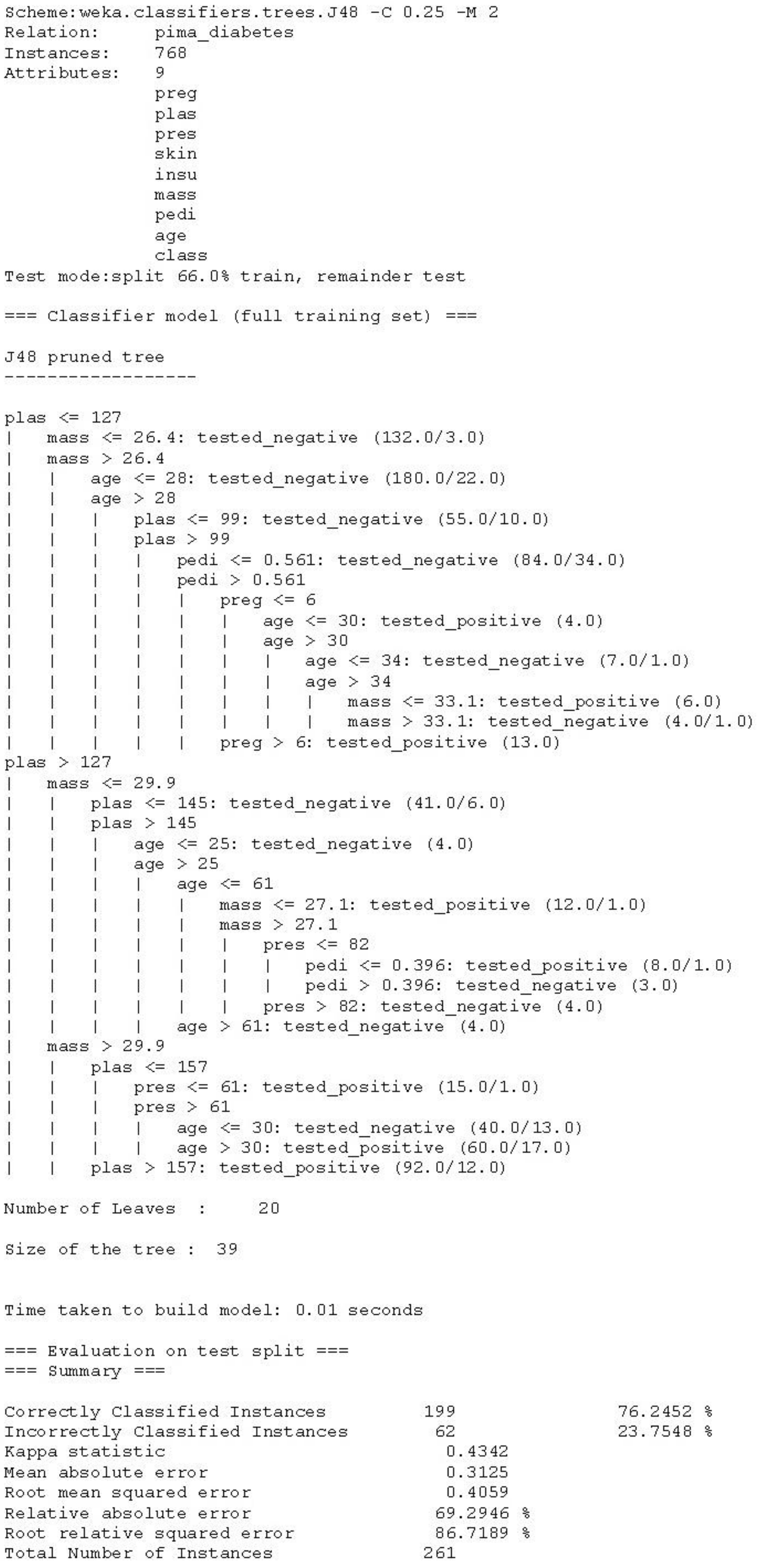
Secondly, we tested whether a decision tree built on encrypted data could be of any use to the data owner and how to make use of it. We will demonstrate the approach with the Pima Indian Diabetes Dataset [68,69]. This dataset has 768 instances and 8 attributes (columns or measurements that describe each instance): number of times pregnant (preg); plasma glucose concentration after 2 hours in an oral glucose tolerance test in mg/dL (plas); diastolic blood pressure in mm Hg (pres); triceps skin fold thickness in mm (skin); 2-hour serum insulin in µU/mL (insu); BMI in kg/m^2 (mass); diabetes pedigree function (pedi); and age in years (age). The final prediction class, actually a rule based on measurements, was tested negative for diabetes (tested_negative) or tested positive for diabetes (tested_positive).

Based on the data, an external analyst (a medical expert) should be able to construct a decision tree that would be able to assist in diagnosing diabetes mellitus for each individual represented by data values in a record (tuple). The decision tree constructed from the original plain text dataset is depicted in Figure 2. The same tree built on encrypted data is depicted in Figure 3.

The trees are identical, except for the attribute names and values, which are encrypted. For example, if the data owner would like to decrypt the encrypted decision rule (see Multimedia Appendix 2), which would read “if [encrypted data]≤255 and [encrypted data]>53.8 and [encrypted data]≤57 then [encrypted answer]”, as seen in lines 1, 3, and 4 from the pruned decision tree, he or she would simply query the lookup table using Structured Query Language (SQL) or any SQL-based graphical tool [70]. The list of queries and the results are shown in Table 4.

Thus, the final decision rule, which was previously encrypted, now reads: IF 2_hr_postload_plasma_glucose <= 127 ∧ body_mass_idx > 26.4 ∧ age <= 28 THEN tested_negative (if 2-hour postload plasma glucose level is ≤127 mg/dL and BMI >26.4 kg/m^2 and age ≤28 years then predict negative diabetes diagnosis).

Figure 2. Decision tree model to assist diagnosing diabetes mellitus built with plain text data from the Pima Indians Diabetes Dataset.



```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     pima_diabetes
Instances:    768
Attributes:   9
              preg
              plas
              pres
              skin
              insu
              mass
              pedi
              age
              class
Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree
------------------

plas <= 127
|   mass <= 26.4: tested_negative (132.0/3.0)
|   mass > 26.4
|   |   age <= 28: tested_negative (180.0/22.0)
|   |   age > 28
|   |   |   plas <= 99: tested_negative (55.0/10.0)
|   |   |   plas > 99
|   |   |   |   pedi <= 0.561: tested_negative (84.0/34.0)
|   |   |   |   pedi > 0.561
|   |   |   |   |   preg <= 6
|   |   |   |   |   |   age <= 30: tested_positive (4.0)
|   |   |   |   |   |   age > 30
|   |   |   |   |   |   |   age <= 34: tested_negative (7.0/1.0)
|   |   |   |   |   |   |   age > 34
|   |   |   |   |   |   |   |   mass <= 33.1: tested_positive (6.0)
|   |   |   |   |   |   |   |   mass > 33.1: tested_negative (4.0/1.0)
|   |   |   |   |   preg > 6: tested_positive (13.0)
plas > 127
|   mass <= 29.9
|   |   plas <= 145: tested_negative (41.0/6.0)
|   |   plas > 145
|   |   |   age <= 25: tested_negative (4.0)
|   |   |   age > 25
|   |   |   |   age <= 61
|   |   |   |   |   mass <= 27.1: tested_positive (12.0/1.0)
|   |   |   |   |   mass > 27.1
|   |   |   |   |   |   pres <= 82
|   |   |   |   |   |   |   pedi <= 0.396: tested_positive (8.0/1.0)
|   |   |   |   |   |   |   pedi > 0.396: tested_negative (3.0)
|   |   |   |   |   |   pres > 82: tested_negative (4.0)
|   |   |   |   age > 61: tested_negative (4.0)
|   mass > 29.9
|   |   plas <= 157
|   |   |   pres <= 61: tested_positive (15.0/1.0)
|   |   |   pres > 61
|   |   |   |   age <= 30: tested_negative (40.0/13.0)
|   |   |   |   age > 30: tested_positive (60.0/17.0)
|   |   plas > 157: tested_positive (92.0/12.0)

Number of Leaves  :     20

Size of the tree :  39


Time taken to build model: 0.01 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         199               76.2452 %
Incorrectly Classified Instances        62               23.7548 %
Kappa statistic                          0.4342
Mean absolute error                      0.3125
Root mean squared error                  0.4059
Relative absolute error                 69.2946 %
Root relative squared error             86.7189 %
Total Number of Instances              261
```

Figure 3. Decision tree model to assist in diagnosing diabetes mellitus built with encrypted data.

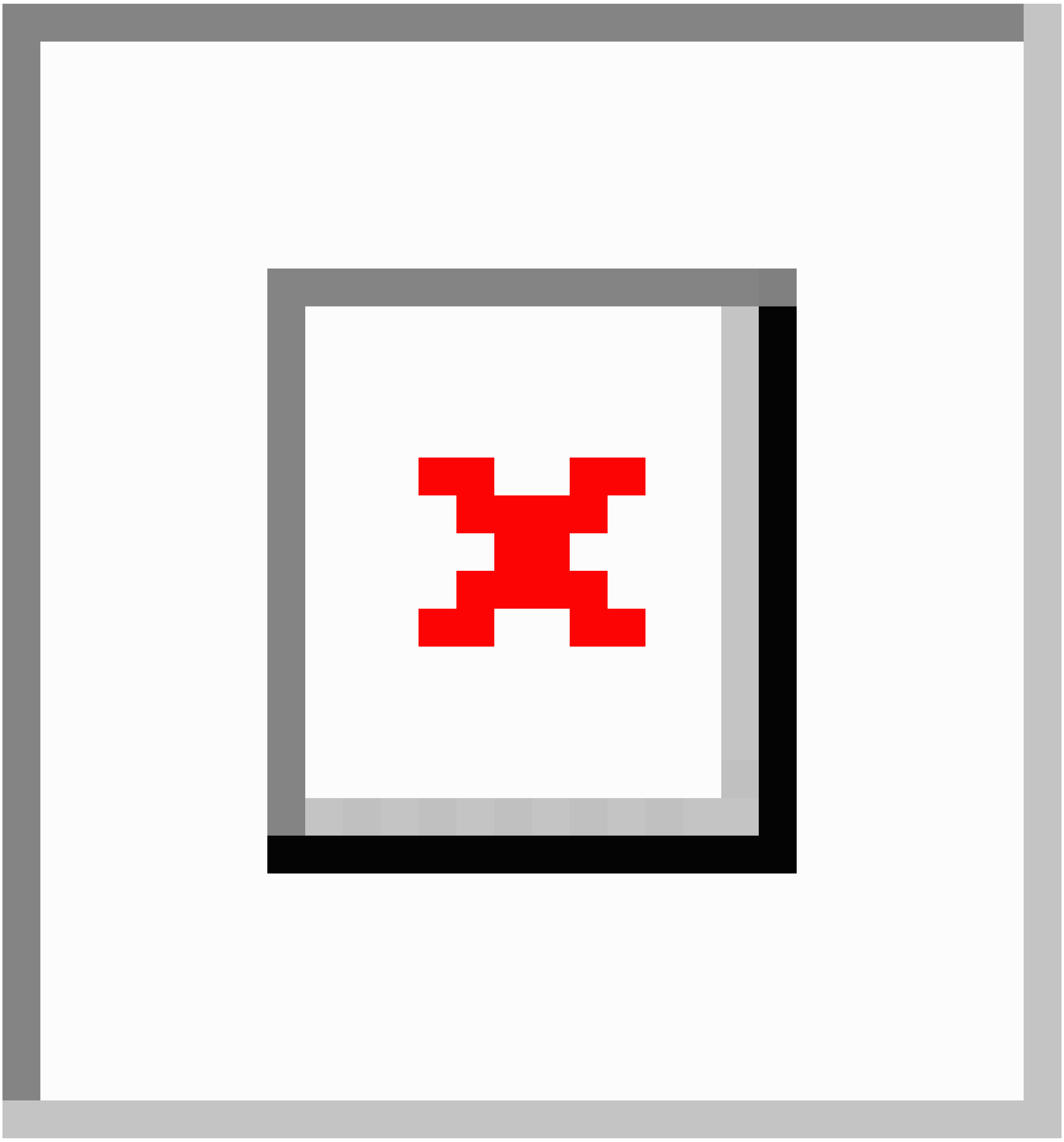


Table 4. List of queries for transforming encrypted data to original.

Query	Result
SELECT original_atribute FROM lookup_table WHERE renamed_attribute=” U2FsdGVkX1/GxsbGxsbGxqJGWHYKVll/Ghr/VuGPcjE=”	2_hr_postload_plasma_glucose
SELECT original_value FROM lookup_table WHERE encrypted_value=255 AND attribute_name=”2_hr_postload_plasma_glucose”	127
SELECT original_atribute FROM lookup_table WHERE renamed_attribute=” U2FsdGVkX1/GxsbGxsbGxnCuqWM8tY1K+ndRFEKNw6w=”	Body mass index
SELECT original_value FROM lookup_table WHERE encrypted_value=53.8 AND attribute_name=”Body mass index”	26.4
SELECT original_atribute FROM lookup_table WHERE renamed_attribute=” U2FsdGVkX1/GxsbGxsbGxht0/44acjje2SK5W1ldQ24=”	Age
SELECT original_value FROM lookup_table WHERE encrypted_value=57 AND attribute_name=”Age”	28
SELECT original_value FROM lookup_table WHERE encrypted_value=” U2FsdGVkX1/GxsbGxsbGxkb6Q6IlZ7BQpR5rsBg4Oi0=” AND attribute_name=”Class”	Tested_negative

Discussion

Principal Findings

This study evaluated the use of a proposed algorithm on 30 publicly available datasets from the UCI repository. The aim of the study was to assess whether analyses on properly protected and encrypted data are possible without a data analyst having access to the original plain text data, which may not always be available because of internal and external restrictions of health care provider institutions.

The analysis showed that the results of building decision trees on original and protected (encrypted) data using the proposed algorithm are virtually identical. The trees were not significantly different based on the bootstrapped *t* tests (*P*=.19).

A decision tree, built on protected data (and the data themselves), is useless for an adversary because all the data are encrypted. The data owner can query the original source data to transform the encrypted data back to readable plain text.

Use Cases and Limitations

There are 3 scenarios or reasons why one might consider using our solution. First, a lack of knowledge and expertise within the health care institution may prevent data processing for decision-making analyses. Secondly, the available resources may not be adequate to perform the analyses. Thirdly, the adherence to an organization’s security policy (eg, based on a need-to-know basis) may not allow for the analyses to be performed on unprotected data. Let us discuss the scenarios in more detail.

The first scenario is when there is a lack of knowledge to do the actual data processing for decision-making purposes and knowledge discovery. In the future, in-house data scientists may be trained so that this becomes less of a practical issue in the business context, but it may remain in the health care environment where the core business is providing health care–related services, not data analyses.

The second scenario is about the lack of computing-related resources that may be available within the health care institution. Although the proposed solution is demonstrated using open-source tools and datasets, customized third-party tools may be needed, which would involve third-party specialists in decision making. When the reason to outsource is because of insufficient computing resources (eg, computing power or storage capacity), one may consider using cloud computing services. These services still need to address the privacy issues (eg, [71,72]), although it is feasible to scale the proposed approach to use multiprocessor power available in the cloud computing environment [73-75].

The third scenario addresses the restrictions imposed on data processing that might exist within the organization because of internal or external rules and regulations. A restriction about data processing could have been set by an internal security policy [76], such as when a security policy based on a need-to-know rule is enforced (eg, [77-79]). The need-to-know rule specifies that the access to sensitive (medical) data is allowed only to those who need to know these data to perform their jobs. Typically, only medical personnel (doctors, nurses) need to have access to specific records to perform their job. Data analysts do not need the access to the undisclosed and unprotected data to perform their job. Rather, they can perform their duties on encrypted data, as suggested by our approach.

All 3 scenarios indicate that the data owner wants to outsource the ability to process the data without actually giving the processor access to it. This may seem contradictory, but we have shown that our approach is feasible if the data are encrypted. In this case, the data processor does not have (nor does she need) the key to decrypt the data, so she does not have the access to the plain text data. The data are safe even when being processed by a third party, or internally when the security policy requires it to be.

Nevertheless, an important limitation should be highlighted. Namely, the plain text data need to be either in numeric or textual (categorical or strings) form. The approach does not support the data mining or decision making on purely textual data. For these to be supported, further work is needed, including the incorporation of a homomorphic encryption scheme [80].

Nonetheless, most of the existing decision-making tools use numeric or categorical data.

Comparison With Prior Work

The approach developed within our study can be used in conjunction with approaches presented by Adam and Wortman [51], Agrawal and Aggarwal [54], or by Agrawal and Srikant [53]. The approaches developed or presented by these authors aim toward blurring or not disclosing the original data. Their approach would produce slightly different analysis results if the original data were used. Our approach can, nevertheless, be applied after one of the previously developed approaches to fully prevent reconstructing the original data by means of statistical disclosure mechanisms.

The proposed solution follows the 7 principles [81] of the PbD [9-11] framework:

1. Proactive not reactive, preventive not remedial: Data are preventively encrypted so any disclosure has no intermediate consequences.
2. Privacy as the default setting: The maximum degree of privacy is delivered by ensuring that personal data are automatically protected in any given IT system or business practice.
3. Privacy embedded into design: Privacy is embedded into the design and architecture of health care IT systems and business practices.
4. Full functionality-positive-sum, not zero-sum: Accommodate all legitimate interests and objectives in a positive-sum win–win manner, not through a dated, zero-sum approach, where unnecessary trade-offs are made. The pretense of false dichotomies, such as privacy vs security, is avoided demonstrating that it is possible to have both.
5. End-to-end security-full lifecycle protection: By having the encryption embedded into the system before the first element of information is stored, the protection is extended throughout the entire lifecycle of the data involved, including during processing.
6. Visibility and transparency (keep it open): The component parts and operations, as proposed by our approach, remain visible and transparent to users and providers alike.
7. Respect for user privacy (keep it user-centric): By using the approach, the architects and operators are required to keep the interests of the individual uppermost by offering such measures as strong privacy defaults.

Conclusions

Medical data stored in online systems are true goldmines. However, if they are not analyzed, they are useless. The problem of data analyses within health care organizations is that these organizations' primary focus is providing health care services and they rarely have enough computing and employee resources to do the analyses. The obvious choice is to use external third-party analysis services. However, exporting sensitive medical data to the outside world can be exposed to significant risks and keeping the medical data safe within health care organizations is also an organizational and technological challenge. Being responsible for someone else's potential mistakes can easily tip the decision toward not using external analyses. Because of time constraints, many health care goals, and the tasks or decisions needed to pursue those goals, these are intentionally deferred until a future opportunity [82], if it ever comes.

It was observed that traditional EHRs have no security guarantee outside the health care service domain [15,16]. The technology that can help is available: the proposed algorithm can be considered as an interface between a data owner from the health care service domain on one side and an outside data analyst on the other side. The design of the algorithm is such that the data are protected in a manner that data analyses are still possible, yet not decipherable by a third party at the same time. Thus, the algorithm conforms to the strict regulations regarding the use and processing of medical data, such as the HIPAA rules and the EU's Directive 95/46/EC. Any potential breach that would involve the data protected with the proposed algorithm is exempt from the HITECH Breach Notification Rule.

In our study, we investigated the feasibility of using encryption within the decision-making process. We tested our approach on 30 databases only. As a part of our future work, further studies with different databases and different types of decisions will be performed to confirm this study's results.

The results of our research confirm that data analyses conducted on protected data can be equivalent to those on original unprotected data. This study's results are promising and provide evidence that the method works. However, more study is needed to show that the method works in all cases.

The procedure can be fully automated. The data owner and the data analyst can seamlessly exchange the data and the results. Importantly, the data and the results are safe while in transit and during processing with the data analyst. The data analyst is not required to implement any additional security measures because these were already implemented at the data owner's side. The proposed approach is compatible with all 7 foundational principles of the PbD framework. By following the PbD framework, we can harness large amounts of data to gain valuable insights into the health system and the health of populations to improve clinical outcomes and achieve cost efficiencies without intruding on privacy [83].

Conflicts of Interest

None declared.

Multimedia Appendix 1

Formal description of data-protecting algorithm.

[PDF File (Adobe PDF File), 128KB-Multimedia Appendix 1]

Multimedia Appendix 2

The algorithm for protection of data for decision-making analyses.

[PDF File (Adobe PDF File), 120KB-Multimedia Appendix 2]

Multimedia Appendix 3

Datasets from UCI Repository in original UCI format and/or ARFF format.

[ZIP File (Zip Archive), 9MB-Multimedia Appendix 3]

Multimedia Appendix 4

Decision trees built on original data.

[ZIP File (Zip Archive), 230KB-Multimedia Appendix 4]

Multimedia Appendix 5

Datasets from UCI Repository in ARFF format, protected with the proposed algorithm.

[ZIP File (Zip Archive), 11MB-Multimedia Appendix 5]

Multimedia Appendix 6

The proposed algorithm prototype with limited functionality implemented in JavaScript.

[ZIP File (Zip Archive), 13KB-Multimedia Appendix 6]

Multimedia Appendix 7

Decision trees built on protected data.

[ZIP File (Zip Archive), 280KB-Multimedia Appendix 7]

Multimedia Appendix 8

SPSS files with result of bootstrapped two dependent samples paired *t* test.

[ZIP File (Zip Archive), 14KB-Multimedia Appendix 8]

References

1. Coldman B. King of the mountain: Digging data for a healthier world. Stanford Medicine Magazine 2012;29(2):20-24 [FREE Full text]
2. Hirschler B. Reuters US Edition. 2010 Dec 01. Roche fears drug industry drowning in "spam" data URL: http://www.reuters.com/article/2010/12/01/roche-data-idUSLDE6B01XF20101201 [accessed 2012-11-12] [WebCite Cache ID 6C7aBng8x]
3. Moore RL, D'Aoust J, McDonald RH. Disktape storage cost models. In: Proceedings of the IS&T Archiving. 2007 Presented at: IS&T Archiving; May 21-25, 2007; Arlington, VA p. 29-32.
4. He W, Zha S, Li L. Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management 2013 Jun;33(3):464-472. [doi: 10.1016/j.ijinfomgt.2013.01.001]
5. Post SG. Encyclopedia of Bioethics. New York: Macmillan Reference USA; 2004.
6. GlobalWebIndex Reports. 2013. Stream Social Global Report, Q1 2013 URL: http://blog.globalwebindex.net/Stream-Social [accessed 2013-11-25] [WebCite Cache ID 6LOPTTrf8]
7. Rooney B. The Wall Street Journal. 2013 Apr 19. Facebook Understands Europe's Privacy Fears Says Sandberg URL: http://blogs.wsj.com/tech-europe/2013/04/19/facebook-understands-europes-privacy-fears-says-sandberg/tab/print/ [accessed 2013-05-23] [WebCite Cache ID 6GpDRK6Fu]
8. Segall L. CNN Money. 2011. Facebook was "the first innovator in privacy" COO says URL: http://money.cnn.com/2011/12/01/technology/facebook_privacy/index.htm [accessed 2013-05-23] [WebCite Cache ID 6GpDuQ7Tn]
9. Cavoukian A. Privacy by Design: Take the Challenge. Toronto, ON: Office of the Privacy Commissioner (Ontario); 2009 Mar 30. URL: http://www.ipc.on.ca/images/Resources/PrivacybyDesignBook.pdf [accessed 2013-12-09] [WebCite Cache ID 6LkJCqPfQ]

10. Cavoukian A, Chanliau M. Privacy and Security by Design: A Convergence of Paradigms. Ontario, Canada: Office of the Privacy Commissioner (Ontario); 2013 Feb 28. URL: http://www.ipc.on.ca/images/resources/pbd-convergenceofparadigms.pdf [accessed 2013-12-09] [WebCite Cache ID 6LkJHxlbH]
11. Cavoukian A, Tapscott D. Who Knows: Safeguarding Your Privacy in a Networked World. New York: McGraw-Hill; 1997.
12. Allen AL. Privacy Law and Society. St Paul, MN: Thomson/West; 2007.
13. Allen AL. The Stanford Encyclopedia of Philosophy (Spring 2011 Edition). 2011. Privacy and Medicine URL: http://plato.stanford.edu/entries/privacy/ [accessed 2012-05-23] [WebCite Cache ID 67s9rEhDU]
14. Privacy Rights Clearinghouse. 2012. Chronology of Data Breaches: Security Breaches 2005-Present URL: http://www.privacyrights.org/sites/privacyrights.org/files/static/Chronology-of-Data-Breaches_-_Privacy-Rights-Clearinghouse.csv [accessed 2012-09-14] [WebCite Cache ID 6AfYySWJK]
15. Anciaux N, Benzine M, Bouganim L. Restoring the patient control over her medical history. 2008 Presented at: 21st IEEE International Symposium on Computer-Based Medical Systems, CBMS '08; June 17-19, 2008; Jyväskylä, Finland p. 132-137. [doi: 10.1109/cbms.2008.101]
16. Ruotsalainen PS, Blobel BG, Seppälä AV, Sorvari HO, Nykänen PA. A conceptual framework and principles for trusted pervasive health. J Med Internet Res 2012 Apr;14(2):e52 [FREE Full text] [doi: 10.2196/jmir.1972] [Medline: 22481297]
17. van der Linden H, Kalra D, Hasman A, Talmon J. Inter-organizational future proof EHR systems. A review of the security and privacy related issues. Int J Med Inform 2009 Mar;78(3):141-160. [doi: 10.1016/j.ijmedinf.2008.06.013] [Medline: 18760661]
18. Solove DJ. A taxonomy of privacy. The University of Pennsylvania Law Review 2006;154(3):477-564.
19. Barnes J. Complete works of Aristotle. In: The Complete Works of Aristotle: The Revised Oxford Translation. Princeton, NJ: Princeton University Press; 1995.
20. Roy J. 'Polis' and 'Oikos' in Classical Athens. Greece & Rome 1999;46(1):1-18 [FREE Full text]
21. Shields JM. A Sacrifice to Athena: Oikos and Polis in Sophoclean Drama. 1991. URL: http://www.facstaff.bucknell.edu/jms089/Z-Unpublished%20Work/Athena.pdf [accessed 2012-09-26] [WebCite Cache ID 6Axq3vRN7]
22. Warren SD, Brandeis LD. The right to privacy. Harvard Law Review 1890;4(5):193-220 [FREE Full text]
23. DeCew JW. The Stanford Encyclopedia of Philosophy. Stanford, CA: Stanford University; 2008. Privacy URL: http://plato.stanford.edu/entries/privacy/ [accessed 2012-05-23] [WebCite Cache ID 67s9rEhDU]
24. Prosser WL. Privacy. California Law Review 1960;48(3):383-423 [FREE Full text]
25. Nagel T. Concealment and Exposure: And Other Essays. Oxford: Oxford University Press; 2002.
26. Moore AD. Privacy: it's meaning and value. American Philosophical Quarterly 2003;40(3):215-227.
27. Gavison R. Privacy and the limits of law. Yale Law Journal 1980;89(3):421-471 [FREE Full text]
28. Allen AL. Uneasy Access: Privacy for Women in a Free Society. Totowa, NJ: Rowman & Littlefield; 1988.
29. Bok S. Secrets: On the Ethics of Concealment and Revelation. New York: Vintage Books; 1983.
30. DeCew JW. The priority of privacy for medical information. Social Philosophy and Policy 2000;17(2):213-234. [doi: 10.1017/S026505250000217X]
31. Allen AL. Genetic privacy: emerging concepts And values. In: Rothstein MA, editor. Genetic Secrets: Protecting Privacy and Confidentiality in the Genetic Era. New Haven, CT: Yale University Press; 1997.
32. Kenny DJ. Confidentiality: the confusion continues. J Med Ethics 1982 Mar;8(1):9-11 [FREE Full text] [Medline: 7069738]
33. Sharpe VA. Privacy and security for electronic health records. Hastings Cent Rep 2005;35(6):49. [Medline: 16396204]
34. United States Senate. Hearings before the Subcommittee on Technology the Law of the Committee on the Judiciary. United States Senate, One Hundred Third Congress, first and second sessions. October 27, 1993, and January 27, 1994. In: High-tech privacy issues in health care: hearings before the Subcommittee on Technology and the Law of the Committee on the Judiciary, United States Senate, One Hundred Third Congress, first and second sessions... October 27, 1993, and January 27, 1994. Washington, DC: US GPO; 1994:141-142.
35. Public Law 104–191: Health Insurance Portability Accountability Act. Washington, DC: US Government; 1996. URL: http://www.gpo.gov/fdsys/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf [accessed 2013-11-02] [WebCite Cache ID 6KpyJGjxy]
36. Department of Health and Human Services. Federal Register. Washington, DC: Office of the Secretary; 2002 Aug 14. Standards for Privacy of Individually Identifiable Health Information; 45 CFR Parts 160 and 164 URL: http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/privrulepd.pdf [accessed 2013-12-10] [WebCite Cache ID 6LlJFBtXT]
37. Department of Health and Human Services. Federal Register. Washington, DC: Office of the Secretary; 2003 Feb 20. Health Insurance Reform: Security Standards; 45 CFR Parts 160, 162 and 164 URL: http://www.hhs.gov/ocr/privacy/hipaa/administrative/securityrule/securityrulepdf.pdf [accessed 2013-12-10] [WebCite Cache ID 6LlJKlyWD]
38. US Department of Health and Human Services. 2012. Summary of the HIPAA Security Rule URL: http://www.hhs.gov/ocr/privacy/hipaa/understanding/srsummary.html [accessed 2012-09-17] [WebCite Cache ID 6AjvRCCgf]
39. Department of Health and Human Services. Federal Register. Washington, DC: Office of the Secretary; 2009 Aug 24. HITECH Breach Notification Interim Final Rule URL: http://www.gpo.gov/fdsys/pkg/FR-2009-08-24/pdf/E9-20169.pdf [accessed 2013-12-10] [WebCite Cache ID 6LlJRW0AM]

40. US Department of Health and Human Services. 2012. Health Information Privacy: HITECH Breach Notification Interim Final Rule URL: http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/breachnotificationifr.html [accessed 2012-05-22] [WebCite Cache ID 67qTJ4Opw]
41. EUR-Lex. 1995 Nov 23. Directive 95/46/EC of the European Parliamentof the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal dataon the free movement of such data URL: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:NOT [accessed 2013-12-10] [WebCite Cache ID 6LlJgqfcX]
42. EUR-Lex. 2001 Jan 12. Regulation (EC) No 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data URL: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32001R0045:en:HTML [accessed 2013-12-10] [WebCite Cache ID 6LlJq5EJR]
43. EUR-Lex. 2012. Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, the free movement of such data URL: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52012PC0010:en:NOT [accessed 2013-05-21] [WebCite Cache ID 6GmQI0E0p]
44. Sengupta S. The New York Times. 2012 Jan 24. Europe weighs tough law on online privacy URL: http://www.nytimes.com/2012/01/24/technology/europe-weighs-a-tough-law-on-online-privacy-and-user-data.html?pagewanted=all&_r=0[WebCite Cache ID 6LlK0Qjdi]
45. Sengupta S. The New York Times. 2012 Jan 24. Facebook's Sandberg gently warns Europe about privacy rules URL: http://bits.blogs.nytimes.com/2012/01/24/facebooks-sandberg-gently-warns-europe-about-privacy-rules/ [accessed 2013-05-21] [WebCite Cache ID 6GmRYxbkO]
46. legislation.gov.uk. 1998. Data Protection Act 1998 URL: http://www.legislation.gov.uk/ukpga/1998/29/contents [accessed 2013-11-01] [WebCite Cache ID 6KoWdlyKS]
47. Bundesministerium der Justiz. 2003. Federal Data Protection Act in the version promulgated on 14 January 2003 (Federal Law Gazette I p. 66), as most recently amended by Article 1 of the Act of 14 August 2009 (Federal Law Gazette I p. 2814) URL: http://www.gesetze-im-internet.de/englisch_bdsg/englisch_bdsg.html [accessed 2013-12-10] [WebCite Cache ID 6LlKPmn4q]
48. NHS Information Governance. 2008 Jan 31. Guidelines on use of encryption to protect person identifiable and sensitive information URL: http://www.connectingforhealth.nhs.uk/systemsandservices/infogov/security/encryptionguide.pdf [accessed 2012-10-08] [WebCite Cache ID 6BG63xlpv]
49. British Medical Association, NHS Connecting for Health. Joint guidance on protecting electronic patient information. 2008. URL: http://www.connectingforhealth.nhs.uk/systemsandservices/infogov/links/jointguidance.pdf/at_download/file [accessed 2012-10-08] [WebCite Cache ID 6BG6cI6pJ]
50. Clifton C, Marks D. Security and privacy implications of data mining. In: Proceedings of the ACM SIGMOD Workshop. New York, NY: ACM; 1996 Presented at: ACM SIGMOD Workshop on Data Mining and Knowledge Discovery; June 1996; Montreal, QC p. 15-19.
51. Adam NR, Worthmann JC. Security-control methods for statistical databases: a comparative study. ACM Comput Surv 1989;21(4):515-556. [doi: 10.1145/76894.76895]
52. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care 2006 Jan;29 Suppl 1:S43-S48. [Medline: 16373932]
53. Agrawal R, Srikant R. Privacy-preserving data mining. SIGMOD Rec 2000 Jun 01;29(2):439-450. [doi: 10.1145/335191.335438]
54. Agrawal D, Aggarwal CC. On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium. New York, NY: ACM; 2001 May 01 Presented at: ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems; May 21-24, 2001; Santa Barbara, CA. [doi: 10.1145/375551.375602]
55. El Emam K, Moreau K, Jonker E. How strong are passwords used to protect personal health information in clinical trials? J Med Internet Res 2011;13(1):e18 [FREE Full text] [doi: 10.2196/jmir.1335] [Medline: 21317106]
56. Dell'Amico M, Michiardi P, Roudier Y. Password strength: An empirical analysis. In: Proceedings of the 2010 IEEE INFOCOM. Piscataway, NJ: IEEE; 2010 Presented at: IEEE INFOCOM; March 14-19, 2010; San Diego, CA p. 1-9. [doi: 10.1109/infcom.2010.5461951]
57. Morris R, Thompson K. Password security: a case history. Commun ACM 1979;22(11):594-597. [doi: 10.1145/359168.359172]
58. Barker WC, Barker E. SP 800-67: Recommendation for the Triple Data Encryption Algorithm (TDEA) Block Cipher. Gaithersburg, MD: National Institute of Standards and Technology; 2012. URL: http://csrc.nist.gov/publications/nistpubs/800-67-Rev1/SP-800-67-Rev1.pdf [accessed 2013-12-10] [WebCite Cache ID 6LlLHuSsN]
59. Federal Information Processing Standards Publication 197: Advanced Encryption Standard (AES). Springfield, VA: National Technical Information Service; 2001. URL: http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf [accessed 2013-12-10] [WebCite Cache ID 6LlLWEbeb]

60. National Institute of Standards and Technology. SKIPJACK and KEA Algorithm Specifications. 1998. URL: http://csrc.nist.gov/groups/STM/cavp/documents/skipjack/skipjack.pdf [accessed 2013-12-10] [WebCite Cache ID 6LlNFPSrS]
61. Stallings W. Cryptography and Network Security: Principles and Practice. Upper Saddle River, NJ: Pearson/Prentice Hall; 2006.
62. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Morgan Kaufman; 2005.
63. Asuncion A, Newman D. UCI Machine Learning Repository. Irvine, CA: University of California at Irvine; 2010. URL: http://archive.ics.uci.edu/ml/datasets.html [accessed 2012-11-09] [WebCite Cache ID 6C2hgsRrX]
64. Asuncion A, Newman D. UCI Machine Learning Repository - selected datasets. Irvine, CA: University of California at Irvine; 2010. URL: http://archive.ics.uci.edu/ml/datasets.html?format=mat&task=cla&att=&area=life&numAtt=&numIns=&type=mvar&sort=nameUp&view=table[WebCite Cache ID 6C2hPCTiA]
65. The Software Environment for the Advancement of Scholarly Research. UCI ARFF Dataset Repository. Champaign, IL: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, SEASR; 2012. URL: http://repository.seasr.org/Datasets/UCI/arff/ [accessed 2012-11-12] [WebCite Cache ID 6C7YkJWal]
66. Quinlan RJ. C4. 5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers; 1993.
67. Cohen PR. Empirical Methods for Artificial Intelligence. Cambridge, MA: MIT Press; 1995.
68. Asuncion A, Newman D. UCI Machine Learning Repository - Pima Indians Diabetes Data Set. Irvine, CA: University of California at Irvine; 2010. URL: http://archive.ics.uci.edu/ml/datasets/Pima%20Indians%20Diabetes [accessed 2012-10-15] [WebCite Cache ID 6BQoPVyou]
69. Smith JW, Everhart J, Dickson W. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Piscataway, NJ: IEEE Computer Society Press; 1988 Presented at: 12th Annual Symposium on Computer Applications and Medical Care; June 1988; Washington, DC.
70. Melton J, Simon AR. SQL: 1999: Understanding Relational Language Components. San Francisco, CA: Morgan Kaufmann; 2002.
71. Kshetri N, Murugesan S. Cloud computing and EU data privacy regulations. Computer 2013 Mar;46(3):86-89. [doi: 10.1109/MC.2013.86]
72. Pearson S, Yee G. Privacy and Security for Cloud Computing. Heidelberg, Germany: Springer; 2013.
73. Dzemyda G, Sakalauskas L. Large-scale data analysis using heuristic methods. Informatica (Lithuan) 2011;22(1):1-10 [FREE Full text]
74. Shotton J, Robertson D, Sharp T. Efficient implementation of decision forests. In: Criminisi A, Shotton JR, editors. Decision Forests for Computer Vision and Medical Image Analysis (Advances in Computer Vision and Pattern Recognition). London: Springer; 2013:313-332.
75. Stahl F, Gaber MM, Bramer M. Scaling Up Data Mining Techniques to Large Datasets Using Parallel and Distributed Processing. London: Springer; 2013:243-259.
76. Pfleeger CP, Pfleeger SL. Security in Computing. Upper Saddle River, NJ: Prentice Hall PTR; 2003.
77. Clark DD, Wilson D. A comparison of commercial and military computer security policies. In: Proceedings of the 1987 IEEE Symposium. Piscataway, NJ: IEEE Computer Society; 1987 Presented at: 1987 IEEE Symposium on Security and Privacy; April 27-29, 1987; Oakland, CA.
78. Lee TMP. Using mandatory integrity to enforce 'commercial' security. In: Proceedings of the IEEE Symposium on Security and Privacy. Piscataway, NJ: IEEE Computer Society; 1988 Presented at: The 1998 IEEE Symposium on Security and Privacy; Apr 18-21, 1988; Oakland, CA. [doi: 10.1109/SECPRI.1988.8106]
79. Nash MJ, Poland KR. Some conundrums concerning separation of duty. In: Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy. Piscataway, NJ: IEEE Computer Society; 1990 Presented at: The 1990 IEEE Computer Society Symposium on Research in Security and Privacy; May 7-9, 1990; Oakland, CA. [doi: 10.1109/RISP.1990.63851]
80. Gentry C. Computing arbitrary functions of encrypted data. Commun ACM 2010 Mar 01;53(3):97. [doi: 10.1145/1666420.1666444]
81. Cavoukian A. Privacy by Design. 2013. 7 Foundational Principles URL: http://www.privacybydesign.ca/index.php/about-pbd/7-foundational-principles/ [accessed 2013-05-24] [WebCite Cache ID 6GqwiKldy]
82. Karsh BT, Weinger MB, Abbott PA, Wears RL. Health information technology: fallacies and sober realities. J Am Med Inform Assoc 2010;17(6):617-623 [FREE Full text] [doi: 10.1136/jamia.2010.005637] [Medline: 20962121]
83. Cavoukian A. Privacy by Design. 2013. Big Data & Privacy Together - It Is Achievable URL: http://www.privacybydesign.ca/index.php/big-data-privacy-together-is-achievable/ [accessed 2013-05-23] [WebCite Cache ID 6GpVPLaXK]

Abbreviations

AES: Advanced Encryption Standard
ARFF: Attribute-Relation File Format
EHR: electronic health records

ePHI: electronic protected health information
HHS: US Department of Health and Human Services
HIPAA: Health Insurance Portability and Accountability Act of 1996
HITECH: Health Information Technology for Economic and Clinical Health Act
KEGG: Kyoto Encyclopedia of Genes and Genomes
NIST: National Institute of Standards and Technology
PbD: Privacy by Design
PHI: protected health information
SEASR: Software Environment for the Advancement of Scholarly Research
SQL: Structured Query Language
TDEA: Triple Data Encryption Algorithm
UCI: University of California at Irvine

Edited by G Eysenbach; submitted 11.12.12; peer-reviewed by P Burnap, N Seeman; comments to author 12.05.13; revised version received 24.05.13; accepted 09.06.13; published 16.12.13

Please cite as:
Brumen B, Heričko M, Sevčnikar A, Završnik J, Hölbl M
Outsourcing Medical Data Analyses: Can Technology Overcome Legal, Privacy, and Confidentiality Issues?
J Med Internet Res 2013;15(12):e283
URL: http://www.jmir.org/2013/12/e283/
doi: 10.2196/jmir.2471
PMID: 24342053