Abstract— Bivariate statistical analysis methods, such as Odds Ratio, Relative Risk, Positive Predictive Value, and Confidence Intervals are very popular analysis methods in every health data research activities. On the other hand, health data is usually gathered and stored by two or more parties, even in different jurisdictions, and because of various privacy acts, they are not allowed or willing to release these data to each other or third parties. In this paper we propose secure procedures for some standard bivariate statistical analysis methods, by which two or more data owners can securely compute those statistics, such that only the final results will be released to the authorized data users. Complexity analysis shows that the cost of applying security features into the standard algorithms is constant and independent from the number of records owned by the data custodians. Also, experimental results provided in this work illustrate a reasonable cost for the proposed secure procedures to prove their applicability on real health research applications.

Keywords— Bivariate Statistical Analysis; Security and Privacy; Health Statistics

Introduction

One of the areas that extensively and intensively use statistical analysis methods is the health research discipline. Health researchers and clinicians apply those analyses for various purposes. For instance, a researcher who wants to investigate the relationship between Aspirin use and Myocardial Infarction, i.e. heart attack, can calculate the Odds Ratio and its confidence interval for the odds ratio for a given $2 \times 2$ contingency table, including the number of people who have used Aspirin and had heart attack or not, and people who have used Placebo and had heart attack or not [1], showed in the following contingency table:

**TABLE I**

<table>
<thead>
<tr>
<th></th>
<th>Heart Attack</th>
<th>No Heart Attack</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>A=189</td>
<td>B=10845</td>
<td>11034</td>
</tr>
<tr>
<td>Aspirin</td>
<td>C=104</td>
<td>D=10933</td>
<td>11037</td>
</tr>
<tr>
<td>Total</td>
<td>293</td>
<td>21778</td>
<td></td>
</tr>
</tbody>
</table>


Computing odds ratio and its confidence interval will help the researcher to find out the effectiveness of taking Placebo in terms of preventing heart attack, comparing to take Aspirin. Using the values in Table I, the estimated odds is 83% higher for those who took placebo compared to Aspirin users.

Another bivariate statistical analysis method is Relative Risk. For instance a clinician is willing to find the
relative risk of lung cancer associated with smoking by using the following $2 \times 2$ contingency table:

**TABLE II**

<table>
<thead>
<tr>
<th></th>
<th>Lung Cancer</th>
<th>No Lung Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>A=30</td>
<td>B=70</td>
<td>100</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>C=10</td>
<td>D=90</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>160</td>
<td></td>
</tr>
</tbody>
</table>

Using the values in Table II, the relative risk of acquiring lung cancer with smoking is 3 times more than that for non-smokers.

By the huge increase of health data availability in various healthcare systems, and the importance of applying accurate data analysis before making decisions and policies from one side, and the importance of preserving the privacy of health data, and data owners’ commitments to truly conform the privacy acts from another side, health researchers and data custodians need to have secure, privacy-preserving techniques and procedures to use for applying various statistical analysis methods on data, which are usually gathered and stored in two or more parties. Many privacy-preserving protocols and algorithms have been provided for data mining, machine learning and statistical analysis methods and algorithms, such as in [2]-[8].

In this paper, we propose privacy-preserving procedures for some popular bivariate statistical analysis methods, namely Odds Ratio, Relative Risk, and their Confidence Interval, Positive Predictive Value, Sensitivity and Specificity. These secure protocols are working for the situations in which data are horizontally partitioned among two or more data owners, such that they are not allowed to reveal their raw data to others, and data users would like to submit their queries, i.e. above-mentioned bivariate statistical analysis methods, and receive only the final results back from the data owners.

In general, there are two different approaches to preserve data privacy. The first one is using data anonymization and perturbation, such as the work proposed in [3], and the second one is Secure Multi-Party Computation (SMC), such as the one in [2]. The main trade-off in the first approach is between the accuracy of the final results and data privacy, and in the former is between efficiency and accuracy of the final results. In this work, because of the importance of the accuracy and reliability of the results, we utilize SMC approach, and the assumption is that the data custodians are correctly following the secure procedures, while they might use the intermediate results to find private data that belong to other parties, i.e. they are semi-trusted parties.

Following is the outline of this paper: Section II discusses background and preliminaries. In Section III we propose the privacy-preserving procedures for the statistical analysis methods. Section IV discusses the security and complexity analysis of the procedures, followed by the experimental results in Section V. Conclusions and future work will be in Section VI.

**Background & Preliminaries**

A. Background

Privacy acts in different jurisdictions prevent data owners to simply reveal and share their health data to other parties. Every health researcher needs to go to a procedure of ethical approval before being able to receive data from data custodians. Even, in many cases, data owners are not willing to disclose their data because of various reasons such as competition, reputation, and numerous policies. To overcome this issue, privacy-preserving techniques and protocols can be proposed and used, such that no raw health data
would be revealed, and at the same time data users can receive aggregated results from the data custodians for their statistical analysis queries.

Since 2000, in which two seminal papers [2]-[3], proposed privacy-preserving data mining techniques using two different approaches, anonymization and secure multi-party computations, there are many protocols and algorithms proposed for various data processing methods. Although randomization and anonymization, using generalization, range, and suppression, have a wide range of usage in many health information centres, these techniques really suffer from utility loss and reliability. On the other hand, SMC procedures, using Private Information Retrieval (PIR) [17]-[19], and Oblivious Transfer techniques [15]-[16], as well as cryptography, especially homomorphic encryption [20]-[24], will not only provide stronger security, it will also guarantee the accuracy of the final results.

In this paper we consider the SMC approach, by using Pallier Cryptosystem [22], which is an RSA-based partially homomorphic encryption. In this cryptosystem for any two plaintext messages $m_1$ and $m_2$, and their encryptions, $E(m_1)$ and $E(m_2)$, we have:

$$D(E(m_1) \times E(m_2)) = m_1 + m_2$$  \hspace{1cm} (1)$$
$$D(E(m_1)^m_2) = m_1 \times m_2$$  \hspace{1cm} (2)

$D$ is decryption operation in the above equations.

Using Pallier cryptosystem, we use Secure Multi-party Multiplication and Secure Multi-party Addition proposed in [8] to implement our secure procedures for the computation of bivariate statistical analysis methods when data is horizontally partitioned among two or more data owners. By horizontally partitioned data, we mean that each data owner has all the information of a subset of records.

Secure Multi-Party Multiplication:

$$\sum_{i=1}^{n} y_i = \prod_{i=1}^{n} x_i$$  \hspace{1cm} (3)

Secure Multi-Party Addition:

$$\prod_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i$$  \hspace{1cm} (4)

In Equations (3) and (4), we assume that $x_i$ is the private input belongs to data owner $i$ and $y_i$ is the private output belongs to data owner $i$.

The main idea is that every data owner will perform some local computations on their data and only encrypted values will be exchanged among them. At the end of the secure procedure, data user will receive some partial information from each data owner and compute the final results by integrating them.

B. Preliminaries

Many statistical and data mining analysis methods are used for various research and healthcare purposes. For instance, adverse drug reactions can be investigated using logistic regression technique [9]-[13]. Another area is using contingency table to study subjects for many health research purposes. For instance in outbreak investigations, in which the categorization is based on exposure of diseases, odds ratio and relative risk and their confidence intervals will be calculated to investigate any possible association between the various exposures and diseases.

For example, researchers want to investigate the association of flea bites and plague using some health data in a specific region and time duration. Following is a 2 X 2 contingency table including synthetic data for this scenario, from epiCentral website [14], that provides resource for epidemiology definitions, examples and activities:

### TABLE III

<table>
<thead>
<tr>
<th></th>
<th>Plague Yes</th>
<th>Plague No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flea Bites Yes</td>
<td>A=40</td>
<td>B=20</td>
<td>60</td>
</tr>
</tbody>
</table>
Now, based on the definition of the bivariate statistical analysis methods, we have the following computations:

A. Odds Ratio (OR)

\[
OR = \frac{A \times D}{B \times C} = \frac{40 \times 30}{20 \times 10} = 6
\]  

(5)

This means that the estimated odds of plague for people who have been exposed to flea bites equals 6 times the estimated odds for people who have not been exposed to flea bites.

B. Relative Risk (RR)

\[
RR = \frac{A/(A+B)}{C/(C+D)} = \frac{40/60}{10/40} = 2.67
\]  

(6)

This means that the people who have been exposed to flea bites are 2.67 times more likely to develop plague disease.

C. Positive Predictive Value (PPV)

\[
PPV = \frac{A}{A+B} = 0.67
\]  

(7)

D. Negative Predictive Value (NPV)

\[
NPV = \frac{D}{C+D} = 0.75
\]  

(8)

E. 95% Confidence Interval of Odds Ratio (CIOR95%)

We should first compute Standard Error (SE) of the log odds ratio using Equation (9) below:

\[
SE(ln(OR)) = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}} = 0.46
\]  

(9)

Then, confidence interval would be computed as follows:

\[
CIOR95\% = \exp\left(\ln(OR) \pm 1.96 \times SE(ln(OR))\right)
\]  

(10)

F. 95% Confidence Interval of Relative Risk (CIRR95%)

We should first compute Standard Error (SE) of the relative risk using Equation (11) below:

\[
SE(ln(RR)) = \sqrt{\frac{1}{A} + \frac{1}{A+B} + \frac{1}{C} + \frac{1}{C+D}} = 0.41
\]  

(11)

Then, confidence interval would be computed as follows:

\[
CIRR95\% = \exp\left(\ln(RR) \pm 1.96 \times SE(ln(RR))\right)
\]  

(12)

G. Sensitivity or True Positive Rate (TPR)

\[
TPR = \frac{A}{A+C} = 0.8
\]  

(13)

H. Specificity or True Negative Rate (TNR)

\[
TNR = \frac{D}{B+D} = 0.6
\]  

(14)

### Secure Bivariate Statistical Analysis Methods

In this section, we propose secure procedures for the bivariate statistical analysis methods mentioned in the previous section. Assumption for data configuration is that there are two or more data owners, each of which maintains a subset of records from the whole data, under which a given data user wants to apply one or more bivariate statistical analysis methods and receive the final results. Here we limit the number of data owners to two, by considering the fact that it can be easily generalized to multi-party case, as well as the space limitation of the paper.

We denote the whole dataset as \( P \) such that \( P = P_1 \cup P_2 \) and \( P_1 \cap P_2 = \emptyset \), and \( P_1 \) and \( P_2 \) are kept by the first and second data owners, respectively. Therefore, for each cell value, say \( A \), of a given \( 2 \times 2 \) contingency table, we will have \( A_1 \in P_1 \) and \( A_2 \in P_2 \) such that \( A = A_1 + A_2 \). Thus, the contingency table would be in the following shape:

**TABLE IV**

**CONTINGENCY TABLE FOR TWO-PARTY HORIZONTALLY PARTITIONED DATA**

<table>
<thead>
<tr>
<th></th>
<th>Disease, Yes</th>
<th>Disease, No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed ( A )</td>
<td>( A_1 + A_2 )</td>
<td>( B_1 + B_2 )</td>
<td>( A+B )</td>
</tr>
<tr>
<td>Non-Exposed ( C )</td>
<td>( C_1 + C_2 )</td>
<td>( D_1 + D_2 )</td>
<td>( C+D )</td>
</tr>
<tr>
<td>Total</td>
<td>( A+C )</td>
<td>( B+D )</td>
<td></td>
</tr>
</tbody>
</table>

Using the distributed contingency table showed in Table IV, following are the proposed secure procedures:

A. Odds Ratio

Distributed version of the Equation (5) would be:
If we expand the fraction in Equation (15), some of them, such as $A_1 \times D_1$ can be calculated locally, as they belong to one data owner. However, some of them, such as $A_1 \times D_2$ should be jointly computed by both parties, without any data privacy violations. For this reason, using secure two-party multiplication, we convert $A_1 \times D_1$ to the summation of two private outputs, such that:

$$A'_1 + D'_2 = A_1 \times D_1$$  \hspace{1cm} (16)

In the Equation (16), $A'_1$ and $D'_2$ are the two private outputs belong to the first and second data owners, respectively. Therefore, Equation (15) will be converted to:

$$OR = \frac{E_1 + E_2}{F_1 + F_2} \hspace{1cm} \text{(17)}$$

such that $E_1$ and $F_1$ are the two private outputs belong to the data owner $i$. Now, by using secure two-party addition, we will have:

$$OR = \frac{E_1 + E_2}{F_1 + F_2} = \frac{G_1 \times G_2}{H_1 \times H_2} = \frac{G_1}{H_1} \times \frac{G_2}{H_2} = K_1 \times K_2$$  \hspace{1cm} (18)

such that $G_i$, $H_i$ and $K_i$ are the three private outputs belong to the data owner $i$. Therefore, to compute odds ratio, the two data owners will apply four secure two-party multiplications and two secure two-party additions using their private inputs and finally each data owner $i$ will send $K_i$ to the data user, whom in turn will compute the final result by multiplying $K_1$ and $K_2$.

### B. Relative Risk

Distributed version of the Equation (6) would be:

$$RR = \frac{(A_1 + A_2) \times (C_1 + C_2 + D_1 + D_2)}{(C_1 + C_2) \times (A_1 + A_2 + B_1 + B_2)}$$  \hspace{1cm} (19)

First note that $C_1 + C_2 + D_1 + D_2$ can be rewritten as $C'_1 + C'_2$ such that $C'_1 = C_1 + D_1$ and $C'_2 = C_2 + D_2$.

Also $A_1 + A_2 + B_1 + B_2$ can be rewritten as $A'_1 + A'_2$ such that $A'_1 = A_1 + B_1$ and $A'_2 = A_2 + B_2$.

Now we should compute the following equation:

$$RR = \frac{(A_1 + A_2) \times (C'_1 + C'_2)}{(C_1 + C_2) \times (A'_1 + A'_2)}$$  \hspace{1cm} (20)

Equation (20) is similar to the equation for odds ratio. Therefore, we apply the same method as for odds ratio to securely compute relative risk.

### C. PPV and NPV

Distributed version of the Equation (7) would be:

$$PPV = \frac{(A_1 + A_2)}{(A_1 + A_2 + B_1 + B_2)}$$  \hspace{1cm} (21)

Using two secure two-party additions, Equation (21) will be converted to multiplication of two private values belong to the data owners as follows:

$$PPV = \frac{(A_1 + A_2)}{(A_1 + A_2 + B_1 + B_2)} = E_1 \times E_2 \hspace{1cm} F_1 \times F_2 = G_1 \times G_2$$  \hspace{1cm} (22)

In Equation (22) $E_i$, $F_1$ and $G_i$ are the three private outputs belong to the data owner $i$, such that $E_1 \times E_2 = A_1 + A_2$ . $F_1 \times F_2 = (A_1 + B_1) + (A_2 + B_2)$, $E_i = G_i$. Therefore, to compute PPV, the two data owners will apply two secure two-party additions using their private inputs and finally each data owner $i$ will send $G_i$ to the data user, whom in turn will compute the final result by multiplying $G_1$ and $G_2$. Secure computation of NPV is similar to computing PPV, and thus we skip its details. Note that secure computation of Sensitivity and Specificity, according to their equations, are similar to the computation of PPV.

### D. Standard Errors of Log Odds Ratio and Relative Risk

Distributed version of the Equation (9) would be:

$$SE(\ln(OR)) = \sqrt{\frac{1}{A_1 + A_2} + \frac{1}{B_1 + B_2} + \frac{1}{C_1 + C_2} + \frac{1}{D_1 + D_2}}$$  \hspace{1cm} (23)

First, we apply secure addition for each fraction to convert each fraction to multiplication of two private outputs belong to the data owners:

$$SE(\ln(OR)) = \sqrt{A'_1 \times A'_2 + B'_1 \times B'_2 + C'_1 \times C'_2 + D'_1 \times D'_2}$$  \hspace{1cm} (24)

Then, by using secure multiplication we convert each multiplication into an addition of two private outputs.
For instance $A_1^2 + A_2^2 = A_1^2 A_2^2$. Finally, Equation (24) will be converted to the square root of an addition of two private outputs, each of which belongs to one data owner. These two private outputs will be securely sent to the data user, whom in turn computes the square root of the addition of those received values to find out the standard error of odds ratio. The same procedure can be applied to securely compute standard error of log relative risk.

**Complexity and Security Analyses**

**A. Complexity Analysis**

As it has been shown in the previous section, two secure building blocks, i.e. secure two-party addition and secure two-party multiplication, are used to securely compute final private shares by the data owners before sending them to the data user. Each of the two building blocks needs two encryptions and one decryption. Cost of the local computations done by each data owner is negligible comparing to the cost of encryptions and decryptions, and can be ignored. Table V illustrates the number of encryptions/decryptions as well as the messages exchanged between the parties, needed for every method.

**TABLE V**

<table>
<thead>
<tr>
<th>Method</th>
<th>Encryption</th>
<th>Decryption</th>
<th>Message Exchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odds Ratio</td>
<td>12</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>Relative Risk</td>
<td>12</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>PPV (NPV)</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Standard Errors</td>
<td>16</td>
<td>8</td>
<td>18</td>
</tr>
</tbody>
</table>

We will see that the number of encryptions/decryptions as well as messages exchanged among the parties are constant and do not depend on the number of dataset records. We have the same situation if we consider more than two data owners, by adding some constants to the values in the Table V.

**B. Security Analysis**

We have assumed that the data owners are semi-honest, i.e. they truly respect all the procedure steps, and provide correct information to each other, while they might use the intermediate results received from each other to reach some private data. As it is usual in SMC approach, we have used Simulation Paradigm [26] and [27], and composition theorem for the proof of procedures security. As we only use secure multi-party addition and multiplication in our procedures, the security proof is exactly the same as the one in [8].

**Experimental Results**

The secure procedures have been implemented using Java programming language, along with the BigInteger class, to be able to efficiently work with very large numbers, as we need to have encryption keys with the minimum length of 1024 bits. Hardware used in the experiment includes Mac OS X, Intel® Core™ i5 2.60GHz, 8 GB DDR3 RAM. Table VI illustrates the performance of each method for a dataset with 100,000 records.

**TABLE VI**

<table>
<thead>
<tr>
<th>Method</th>
<th>Time (in Seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odds Ratio</td>
<td>0.0197</td>
</tr>
<tr>
<td>Relative Risk</td>
<td>0.0197</td>
</tr>
<tr>
<td>PPV (NPV)</td>
<td>0.0071</td>
</tr>
<tr>
<td>Standard Errors</td>
<td>0.0268</td>
</tr>
</tbody>
</table>

**Conclusions & Future Work**

In this paper, we have proposed secure procedures for computing some popular bivariate statistical analysis
methods that are widely used, especially in various health research disciplines. Our assumption is that health data are horizontally partitioned and stored by two or more data owners, and they are willing to jointly answer the queries submitted by a given data user, e.g., a health researcher. Secure computation of multivariate analysis would be a future work to continue the approach used in this paper, along with providing a framework for all these statistical analysis methods in an integrated system, used by health researchers and clinicians.

References


