Current HIV Research, 2016, *14*, 121-129

Applying Triple-Matrix Masking for Privacy Preserving Data Collection and Sharing in HIV Studies

Qinglin Pei[1], Shigang Chen[2], Yao Xiao[3] and Samuel S. Wu[*,1]

[1]*Department of Biostatistics, University of Florida, Gainesville, USA;* [2]*Department of Computer & Information Science & Engineering, University of Florida, Gainesville, USA;* [3]*Institute of National Accounts, Beijing Normal University, Beijing, 100875, P.R. China*

Samuel S. Wu

Abstract: *Background*: Many HIV research projects are plagued by the high missing rate of self-reported information during data collection. Also, due to the sensitive nature of the HIV research data, privacy protection is always a concern for data sharing in HIV studies.

Methods: This paper applies a data masking approach, called triple-matrix masking [1], to the context of HIV research for ensuring privacy protection during the process of data collection and data sharing.

Results: Using a set of generated HIV patient data, we show step by step how the data are randomly transformed (masked) before leaving the patients' individual data collection device (which ensures that nobody sees the actual data) and how the masked data are further transformed by a masking service provider and a data collector. We demonstrate that the masked data retain statistical utility of the original data, yielding the exactly same inference results in the planned logistic regression on the effect of age on the adherence to antiretroviral therapy and in the Cox proportional hazard model for the age effect on time to viral load suppression.

Conclusion: Privacy-preserving data collection method may help resolve the privacy protection issue in HIV research. The individual sensitive data can be completely hidden while the same inference results can still be obtained from the masked data, with the use of common statistical analysis methods.

Keywords: Contingency table analysis, Cox regression, general linear model, logistic regression, privacy-preserving data collection.

1. INTRODUCTION

Since human immunodeficiency virus (HIV) was first clinically observed in 1981, the number of infections has steadily increased over the past three decades. It was estimated by CDC that the HIV incidence in the United States was around 50,000 per year. As of 2010, there were about 1,144,500 persons aged 13 years or older living with HIV infection in the United States, among which about 15.8% were not aware of their infections [2]. With new treatments, approximately 85% of HIV patients survive more than 4 years after the infection [3]. The mandatory domestic care and treatment programs result in significant increment of the federal funding for HIV over the course of the epidemic. The federal budget for HIV in the fiscal year 2015 is estimated to be $30.4 billion, which reflects an increment of $5.4 billion since 2009 [4]. HIV research plays an irreplaceable role in providing better patient care and more reasonable medical resource allocation.

The availability of patient data, particularly truthful data, is critical in HIV studies. However, such data are sensitive to the patients. Studies have shown that disclosing sensitive personal information may lead to risks of discrimination, social stigma, and physical harm, where sensitive information can be related to genetics, mental health, reproductive care, substance abuse and sexually transmitted diseases [5]. In extreme cases, the risks can extend beyond the individual to his or her family, employer or others.

As a result, privacy concern presents a serious obstacle to researchers in their effort to obtain truthful patient data. More specifically, without ensuring privacy during the data collection process, researchers face the problem of missing data in two categories: missing of subjects and missing of items. Missing of subjects refers to those patients who do not consent to release their information in HIV studies. Missing of items refers to those patients who agree to participate in the study but refuse to provide some sensitive self-reported information because they do not have enough trust in the confidentiality protection. A significant amount of HIV data such as substance use, medication adherence and sexual behaviors were often determined by self-reported data and participants often under-report this kind of sensitive information [6]. The importance of HIV self-reported data has been widely acknowledged [7, 8], but the integrity of sexual behavior data has also been questioned since Kinsey's pioneering survey sexuality in 1950's [9, 10]. The rate of missing data in HIV research can be high. For example, in a research studying the relationship characteristics associated with sexual risk behavior among MSM (men who have sex with men) in committed relationships, the Sexual Agreement Investment Scale score had roughly 20% of values missing, which is a nontrivial amount of incomplete data [11]. Such

*Address correspondence to this author at the Department of Biostatistics, University of Florida, Gainesville, USA; Tel: ++01-352-294-5910; E-mail: samwu@biostat.ufl.edu

1873-4251/16 $58.00+.00

missing data can lead to severe underreport and inaccurate estimation.

The relatively high rate of missing or skewed information during data collection can potentially cause bias in HIV research. To address the privacy issue, this paper adapts a generic privacy-preserving data collection method, called triple-matrix masking and proposed by Wu *et al*. (2014) [1] to the context of HIV studies with domain-specific considerations in an effort to relieve the privacy concern and encourage patients to report truthful data in HIV research. The proposed approach enhances current practice with new technologies for full privacy protection, ensuring that raw data stay with patients and only masked data are collected. Not only will de-identification be now performed directly by patient devices right after data are produced, but also the data themselves are completely masked right away, such that sensitive information can be transferred without worry of identification or data leak. These technologies hold the promise of removing the trust obstacle, promoting objective data collection, and helping unrestricted sharing of data across different HIV studies. The masked data do not give out individual information but the statistical inference on parameters of interest can be conducted with the same results on masked data as on the raw data, under general linear model, chi-square test, logistic regression, and Cox proportional hazard regression. The motive of this work is to contribute to the security of sensitive data of HIV patients, beyond the simple removal of personal identifiers from databases. As an added value, this additional security may lead to a less cumbersome IRB approval process and it will encourage data retention and sharing even after research projects are completed.

The rest of the paper is organized as follows: Section 2 discusses the related work. Section 3 briefly summarizes the generic triple-masking method for privacy-preserving data collection. Section 4 adapts the method to the domain of HIV studies through an example, and shows how various statistical analyses can be performed on the masked data. Section 5 draws the conclusion.

2. RELATED WORK

Traditionally, to mitigate the privacy concern in HIV studies, the researchers first try to build patients' confidence in the pre-data collection stage and then apply some remedy methods to handle missing values in the post-data collection stage. In a study to investigate the young age effect on the antiretroviral adherence and viral load suppression among the injection drug users, researchers put a significant effort to assure confidentiality and to build trust with all the participants [12]. Another study was conducted in a predominantly Hispanic-serving community health center in a high HIV prevalence area to understand patient beliefs of who should be tested for HIV in the routine HIV testing era. The researchers reformatted the survey questions: they queried participants on what populations they thought were at risk for HIV and should be tested instead of asking them about their own risks [13]. Despite of all these efforts, self-reported data still carry a large portion of missing or skewed information because the patients know that as long as the raw data are collected, their privacy cannot be fully protected, subject to intentional or unintentional leaks.

In the post-data collection stage, there are many methods to treat missing values. One simple approach is to separate patients with missing values [14]. Other approaches may estimate the missing values by mean substitution, hot deck imputation or regression substitution [15, 16]. Although these methods can mitigate the problem, they cannot completely solve it.

Besides the missing values in the data collection phase, information confidentiality is also a critical issue in the data sharing phase, particular in the era of big data. Various types of HIV patient data such as demographics, clinical measurements, social characteristics and genomic information are collected and cumulated by different data centers. While the data sharing among data centers may lead to broader and more profound results, it faces the legal issue of privacy protection [17-19].

Most existing privacy protection methods are designed to share obfuscated data by entities that have already collected raw data, and they include addition of noise [20-22], multiple imputation [23], information preserving statistical obfuscation [24], post randomization method [25], controlled tabular adjustment [26], data shuffling [27], random projection based perturbation [28], and random orthogonal matrix masking [29]. These methods perform data obfuscation after raw data are collected. They are ineffectiveness against the security breach of the data management centers themselves, which face real threats from the cyberspace, as is evident from the recent well-publicized online stealing of credit card information from major retailers and hacks against bank servers.

3. PRIVACY-PRESERVING DATA COLLEECTION THROUGH TRIPLE-MATRIX MASKING

A new method called triple-matrix masking is developed by Wu *et al*. [1]. It advances the data masking from centralized data centers all the way to patients themselves in order to achieve a full privacy protection. Specifically, the data are randomly transformed at the time of collection to ensure that no raw data could be seen in the entire data lifecycle. However, because of the specially designed transformations, statistical inference on parameters of interest can be conducted with the same results as if the original data were used. This is achieved by assigning multiple keys to different parties.

We give a brief overview of the generic method: The system consists of data providers, data collectors, data users, and a masking service provider. In most applications, data providers are patient participants, and data users are study investigators as well as other researchers who can access the information. Typically, the data managers and statistical analysts in the study investigative team are in charge of data collection. Also, they release transformed data to the data users once the data have been collected. The masking service provider may be a private business or a government entity established to promote data sharing. A masking service provider only receives masked data from data providers and then applies a second mask. The data collectors who hold the key to the first mask partially decrypt the doubly masked

data and apply a third mask before releasing the data to the public. The critical feature of the method is that the keys used to generate the masking matrices are held separately by the masking service provider and the data collectors. This ensures that nobody sees the actual data, but statistical inference on parameters of interest can be conducted with the same results on masked data as on the raw data.

In the following, we adapt the above generic privacy-preserving data collection method to the context of HIV studies with domain-specific considerations in order to address the practically important privacy protection issue in HIV research.

4. PROTECT PRIVACY THROUGH TWO METHODS OF PRIVACY-PRESERVING DATA COLLECTION

4.1. Data Set

We generate patient data based on Hadland *et al*. [12] to study the effect of age on adherence to antiretroviral therapy for HIV and on time to viral load suppression among young injection drug users. The data set contains sensitive information on 7 sociodemographic factors and 8 substance use behaviors, in addition to age, baseline CD4+ cell count and baseline plasma viral load. For simplicity, we present the data masking procedure and how the masked data enable the same statistical inference as the original data using only 30 observations that contain 2 sensitive variables: sex trade involvement and daily heroin use in the last six months (see Table **1** for the generated patient data).

4.2. Orthogonally Record-Transformed Data Preserve Useful Statistics

First we show that, with the use of general linear model and contingency table analysis, orthogonally record-transformed data preserve sufficient statistics and enable us to obtain the exact same analytical results with masked data as with the original data.

Step 1. Before the data collection, the data collector creates a database consisting of the nine variables listed in Table **1**, plus a binary variable for young age (age <=29), a variable for quality assurance and a variable for noise. Also, a web-based data entry system is developed for each participant to enter the data. In addition, the data collector chooses a key of 535 as the random seed to generate a 12 by 12 random invertible matrix B, which is given below.

Step 2. At the time of data collection, the first participant enters its data which are shown in the first row of Table **1**. The record is immediately transformed by B and only the masked data xB, are sent to the masking service provider. This is repeated for all 30 subjects, resulting attribute-transformed data XB.

Step 3. The masking service provider chooses a different key 536, and uses the Matlab program described in the Appendix 1 of Wu *et al*. [1] to generate a 30 by 30 random orthogonal matrix A_2 = GenerateROM(536, 30). Due to space limit, we omit the A_2 matrix here but readers can easily get the matrix by running the Matlab program. After receiving attribute-transformed data from all participants, the masking service provider applies record transformation to XB and sends the doubly masked data, A_2XB, to the data collector.

Step 4. The data collector chooses another key 537 to produce a 30 by 30 random orthogonal matrix A_1= GenerateROM(537, 30), which is generated in a similar way as A_2. Then A_2XB is multiplied by B^{-1} to get back A_2X, which is further left-multiplied by A_1 before publishing. At the end, the data users have access to orthogonally transformed data A_1A_2X (see Table **2**).

Table **2** shows that the transformed data for the binary variables (ADH, Male, STI, DHU, and Young) take real values. From these masked data, users can only guess each participant's sensitive characteristics - whether she or he had sex trade involvement and daily heroin use in the last six months. However, for statistical inference, users can obtain the exact counts for contingency tables based on the masked data A_1A_2X. Specifically, the frequency counts can be obtained from the masked data use the inner-product of the masked vectors. For example, the number of subjects that adhered to antiretroviral therapy (n=8) can be calculated using the sum of squares over the masked ADH variable; and the number of male subjects that adhered (n=4) can be calculated using the sum of cross product of the masked ADH and Male variables. These two numbers, along with the

$$B = \begin{bmatrix} 0.36 & 0.13 & 0.53 & 0.52 & 0.61 & 0.83 & 0.09 & 0.88 & 0.11 & 0.04 & 0.89 & 0.96 \\ 0.75 & 0.19 & 0.65 & 0.78 & 0.09 & 0.89 & 0.32 & 0.21 & 0.56 & 0.11 & 0.73 & 0.72 \\ 0.16 & 0.44 & 0.97 & 0.63 & 0.16 & 0.80 & 0.86 & 0.90 & 0.84 & 0.16 & 0.76 & 0.56 \\ 0.67 & 0.42 & 0.64 & 0.10 & 0.16 & 0.97 & 0.99 & 0.56 & 0.94 & 0.99 & 0.34 & 0.28 \\ 0.67 & 0.69 & 0.50 & 0.19 & 0.32 & 0.82 & 0.63 & 0.97 & 0.10 & 0.03 & 0.69 & 0.74 \\ 0.44 & 0.95 & 0.72 & 0.33 & 0.69 & 0.90 & 0.39 & 0.45 & 0.49 & 0.69 & 0.20 & 0.86 \\ 0.34 & 0.55 & 0.54 & 0.02 & 0.95 & 0.63 & 0.17 & 0.72 & 0.55 & 0.20 & 0.58 & 0.42 \\ 0.48 & 0.11 & 0.49 & 0.64 & 0.47 & 0.87 & 0.06 & 0.21 & 0.76 & 0.17 & 0.65 & 0.98 \\ 0.81 & 0.97 & 0.63 & 0.99 & 0.35 & 0.40 & 0.82 & 0.34 & 0.17 & 0.63 & 0.28 & 0.43 \\ 0.53 & 0.66 & 0.25 & 0.13 & 0.83 & 0.06 & 0.36 & 0.31 & 0.75 & 0.59 & 0.24 & 0.19 \\ 0.55 & 0.76 & 0.16 & 0.70 & 0.06 & 0.36 & 0.54 & 0.82 & 0.74 & 0.46 & 0.07 & 0.10 \\ 0.18 & 0.65 & 0.42 & 0.89 & 0.63 & 0.57 & 0.91 & 0.82 & 0.98 & 0.15 & 0.53 & 0.72 \end{bmatrix}$$

xB = [877.8 1002.7 525.0 554.8 310.7 910.5 844.5 1275.2 523.8 311.6 623.9 682.9],

Table 1. Original data set X, which was generated according to Hadland *et al.* (2012), along with a quality assurance and noise variable.

Time	Censoring	ADH	Age	CD4	Log10PVL	Male	STI	DHU	Young	QA	Noise
12.13	1	0	28	815	3.93	1	0	0	1	555	0.1731
12.71	1	1	37	591	3.76	1	1	1	0	555	0.0087
21.35	1	0	40	715	3.75	1	1	1	0	555	0.9313
3.76	1	0	28	430	4.47	1	1	0	1	555	0.8825
6.32	0	0	44	643	3.65	0	0	0	0	555	0.8402
10.01	1	1	47	283	4.45	0	0	0	0	555	0.1311
1.73	1	1	33	261	4.70	1	0	1	0	555	0.7811
1.52	1	0	54	377	3.33	1	0	0	0	555	0.9929
0.56	1	0	30	133	3.95	1	0	0	0	555	0.8086
2.52	1	0	35	456	3.03	1	1	0	0	555	0.0114
10.05	1	0	25	546	4.44	1	0	0	1	555	0.7029
7.29	1	1	44	424	3.47	0	0	0	0	555	0.9023
2.10	1	0	50	292	5.48	1	1	1	0	555	0.3907
19.91	1	0	49	751	2.63	1	1	1	0	555	0.9351
2.39	0	1	35	573	4.11	1	0	0	0	555	0.0947
0.50	1	0	28	488	4.82	0	0	1	1	555	0.0128
0.66	0	0	33	313	4.21	0	0	0	0	555	0.391
1.06	0	0	46	477	3.71	0	0	1	0	555	0.5762
22.04	1	0	50	293	3.05	1	0	1	0	555	0.5132
1.99	0	1	27	218	4.61	1	0	0	1	555	0.26
1.65	1	0	28	347	4.34	0	1	0	1	555	0.7415
3.80	0	0	32	113	4.87	0	1	0	0	555	0.6293
7.29	0	0	35	114	4.06	1	0	0	0	555	0.8766
0.50	1	1	33	434	4.93	0	0	1	0	555	0.4116
9.31	1	0	30	192	3.73	1	0	0	0	555	0.8075
23.13	0	0	31	372	2.49	0	0	0	0	555	0.5768
2.32	1	0	30	0	6.25	1	1	0	0	555	0.6599
4.60	0	0	44	429	5.30	1	1	0	0	555	0.104
3.32	1	0	42	154	4.80	0	0	0	0	555	0.8714
7.58	0	1	29	626	3.70	0	0	0	1	555	0.0957

Time=Time to viral load suppression (< 500 copies per milliliter); **Censoring**=Censoring indicator (1=not censored, 0= time censored by end of follow-up); **ADH=**Adherence to antiretroviral therapy (Yes=1/No=0); **Age=**Baseline age in years; **CD4=** Baseline CD4+ cell count as a continuous variable in cells/μL; **Log10PVL=**Baseline plasma viral load in log10[number of HIV RNA copies per milliliter]; **Male=**Male gender; **STI=**Sex trade involvement in the last six months (Yes=1/No=0); **DHU=**Daily heroin use in the last six months (Yes=1/No=0); **QA=**Quality assurance.

total sample size, are sufficient to create the 2 by 2 contingency table between ADH and Male; hence the masked data yield exactly the same association analysis results between the two variables (see Table 3). In addition, it has been shown in [1] that any multivariate regression analysis (e.g., Log10PVL predicted by age and CD4 counts) based on the masked data will also give the same results as if the original data are used.

4.3. Attribute-Transformed Data Preserve Useful Statistics

Many applications conduct logistic regression for binary outcomes and Cox regression for survival times. For example, to adjust for gender, STI, DHU, and baseline CD4 and Log10PVL, we may use logistic regression to study the association between age and adherence and use Cox regression to study the association between age and time to

viral suppression. In this case, we reverse the usage of random matrix masking: the data collector generates the row operator A and the masking service provider applies the column operator B. Both operators are invertible matrices, but not required to be orthogonal. The new procedure is as follows:

Step 1. The data collector creates the database structure, programs the data collection system, and chooses a key of 535 to generate an 8 by 8 random invertible matrix A, which is distributed to the participants' data collection devices.

Step 2. At the time of data collection, the first participant's data are independently augmented with 6 extra rows of normal random noise (which the data collector does not know) and a row of quality assurance data (see Table 4a). The augmented data matrix, denoted by x*, is immediately masked and only the transformed data Ax* (see Table 4b) are sent by the participant to the masking service provider.

Step 3. The masking service provider generates a column operator which is constructed to be block-diagonal so that it keeps the first 4 columns invariant and the lower 5 by 5 block is randomly generated. Then the masking service provider applies attribute-transformation B_1, and sends the doubly masked data Ax^*B_1 (see Table 4c) to the data collector.

Step 4. The data collector left-multiplies Ax^*B_1 by inverse of matrix A to get back x^*B_1, and extracts the first row of x^*B_1 to get back xB_1.

Step 5. After receiving such attribute-transformed data from all participants (repeat Steps 1-4), the data are aggregated to be XB_1. Then, the data collector generates a random column operator B_2.

Step 6. The data collector right-multiplies XB_1 by B_2, and publishes XB_1B_2 so that data users have access to the column transformed data.

It is easy to check that the masked data XB_1B_2 provides exactly the same results as the original data X in terms of the planned logistic regression and Cox regression on the effect of age. Specifically, in the logistic regression to study the

$$A=\begin{bmatrix}
0.3622 & 0.8146 & 0.6877 & 0.5300 & 0.6252 & 0.1891 & 0.6139 & 0.3486\\
0.7470 & 0.5330 & 0.9458 & 0.6486 & 0.2512 & 0.3250 & 0.0904 & 0.8303\\
0.1635 & 0.5532 & 0.5465 & 0.9722 & 0.1597 & 0.0221 & 0.1620 & 0.0578\\
0.6691 & 0.1752 & 0.1052 & 0.6382 & 0.4226 & 0.6365 & 0.1629 & 0.6275\\
0.6674 & 0.1261 & 0.9745 & 0.5047 & 0.5198 & 0.9869 & 0.3162 & 0.8318\\
0.4392 & 0.1946 & 0.6600 & 0.7202 & 0.7759 & 0.1257 & 0.6940 & 0.8877\\
0.3429 & 0.4399 & 0.7629 & 0.5385 & 0.6283 & 0.6993 & 0.9477 & 0.8043\\
0.4811 & 0.4247 & 0.6468 & 0.4894 & 0.1014 & 0.8917 & 0.4742 & 0.9711
\end{bmatrix}$$

$$B_1=\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & -0.0417 & -0.2300 & -0.0966 & 0.6893 & 0.6789\\
0 & 0 & 0 & 0 & 0.8559 & -0.1341 & 0.2334 & -0.2890 & 0.3338\\
0 & 0 & 0 & 0 & 0.4556 & 0.3547 & 0.0450 & 0.6442 & -0.4995\\
0 & 0 & 0 & 0 & -0.0309 & 0.8880 & -0.1384 & -0.1348 & 0.4161\\
0 & 0 & 0 & 0 & -0.2389 & 0.1213 & 0.9566 & 0.0903 & 0.0708
\end{bmatrix}$$

$$B_2=\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0.7254 & 0.0021 & -0.3332 & 0.0034 & 0.6023\\
0 & 0 & 0 & 0 & 0.3211 & 0.7867 & 0.3838 & -0.3162 & -0.1754\\
0 & 0 & 0 & 0 & -0.1887 & 0.5267 & -0.2257 & 0.7917 & 0.0960\\
0 & 0 & 0 & 0 & -0.3320 & 0.0064 & 0.5971 & -0.0017 & 0.7302\\
0 & 0 & 0 & 0 & 0.4742 & -0.3219 & 0.5781 & 0.5227 & -0.2531
\end{bmatrix}$$

Table 2. Matrix-masked data released to data users, A_1A_2X.

Time	Censoring	ADH	Age	CD4	Log10PVL	Male	STI	DHU	Young	QA	Noise
2.10	0.82	0.16	36.81	651.16	4.08	0.39	0.64	0.06	0.64	555	-0.0928
6.74	1.19	0.22	29.47	698.86	5.33	1.05	1.22	0.89	0.02	555	0.1781
4.43	0.65	0.45	44.25	112.85	4.92	0.30	0.38	-0.32	-0.07	555	1.0048
13.99	0.87	0.28	59.87	723.31	3.51	0.52	0.69	0.08	-0.35	555	0.777
3.34	1.03	-0.37	40.87	392.55	2.89	1.11	0.35	-0.08	0.06	555	0.9925
7.92	1.13	0.10	38.41	276.44	4.63	0.87	0.83	0.34	0.39	555	0.812
6.82	0.45	0.02	50.21	313.06	4.53	0.69	0.44	0.25	-0.21	555	0.5471
3.70	0.81	0.77	27.13	133.90	4.00	0.99	-0.50	-0.24	0.30	555	0.955
23.71	-0.24	-0.24	33.40	820.82	2.01	1.20	0.58	0.40	0.54	555	0.555
20.27	0.84	0.71	38.88	563.13	3.20	0.35	0.06	0.66	0.55	555	-0.0736
-5.05	0.67	0.44	27.55	478.66	4.95	0.54	1.18	0.26	0.95	555	0.6835
12.04	1.26	0.95	47.76	631.01	3.36	0.47	0.09	1.44	0.18	555	0.7957
19.14	0.71	0.73	36.55	162.74	3.50	1.09	0.84	0.39	-0.41	555	0.6246
3.50	0.49	0.43	24.97	298.59	4.56	0.99	0.62	-0.17	0.28	555	0.5491
12.45	0.36	-0.49	22.01	293.20	4.63	-0.60	-0.20	-0.25	0.93	555	0.8924
7.58	0.36	0.79	22.40	524.95	4.19	0.37	0.32	-0.80	0.93	555	0.2048
0.14	0.31	0.20	35.50	445.83	4.52	1.20	0.28	0.31	0.96	555	0.3654
12.35	0.80	-0.14	41.36	606.04	2.91	0.10	-0.11	0.67	-0.14	555	0.7267
4.73	0.87	0.23	35.89	328.11	4.51	0.63	0.73	0.97	0.11	555	0.3877
6.41	1.42	-0.26	35.74	124.57	5.01	1.82	0.33	0.10	0.08	555	0.3935
3.81	-0.11	0.92	38.46	396.74	4.68	-0.06	0.41	0.68	-0.09	555	0.2429
4.39	0.90	-0.14	31.45	245.04	4.09	0.29	0.68	0.52	0.22	555	0.7566
2.14	0.55	0.26	41.09	96.84	3.20	-0.18	0.37	0.06	-0.38	555	0.2243
8.66	0.81	0.09	35.53	409.02	4.59	0.94	0.46	0.36	0.81	555	0.1285
5.79	1.78	0.20	30.17	522.22	3.71	0.39	-0.04	0.45	0.72	555	0.881
-1.19	-0.39	-0.14	43.26	303.39	3.90	0.11	-0.26	-0.08	-0.09	555	0.6556
1.97	0.74	1.18	39.15	461.83	4.54	0.52	-0.69	0.85	0.42	555	0.2262
-0.96	-0.14	1.00	27.35	435.59	4.20	0.36	-0.69	-0.15	0.00	555	0.094
-2.91	0.42	-0.08	37.62	10.03	6.10	0.41	0.65	0.61	0.02	555	0.5082
16.09	0.65	-0.28	43.91	399.51	3.80	1.13	0.35	0.75	-0.37	555	1.118

*Variable abbreviations are the same as those in Table **1**.

Table 3. Association between adherence to antiretroviral therapy and sociodemographic characteristics as well as recent substance use behaviors.

Variable	Total	Adherence		p-Value
		No (n=22)	Yes (n=8)	
Male gender	18	14 (63.6%)	4 (50.0%)	0.50
Sex trade involvement	10	9 (40.9%)	1 (12.5%)	0.14
Daily heroin use	9	6 (27.3%)	3 (37.5%)	0.59
Young (age<30)	7	5 (22.7%)	2 (25.0%)	0.90

Table 4. Augmented data x*, initially masked data Ax*, and doubly masked data Ax* B_1 for the first observation.

4a. Augmented data x* containing the first observation (first row) along with 6 rows of normal random noise and 1 row for quality assurance.

Time	Censoring	ADH	age	CD4	Log10PVL	Male	STI	DHU
12.13	**1**	**0**	**28**	**815**	**3.93**	**1**	**0**	**0**
-0.73	-0.65	-1.52	0.10	0.17	0.18	0.02	-1.92	0.1439
0.43	-0.07	-1.34	-0.97	0.18	-0.07	1.12	0.19	-0.667
-0.65	0.42	-0.20	1.84	0.96	0.44	0.25	0.86	-0.7951
0.30	0.13	-0.41	0.52	0.33	-0.37	-0.80	-0.98	-1.4992
0.56	0.26	0.63	0.49	0.02	-1.15	-1.59	0.32	-0.7516
-0.23	-0.59	0.94	0.11	0.33	-0.14	-0.70	1.74	0.2906
777	777	777	777	777	777	777	777	777

4b. Initially masked data Ax* for the first observation.

Time	Censoring	ADH	Age	CD4	Log10PVL	Male	STI	DHU
274.78	270.64	269.05	281.89	567.21	272.09	270.93	270.41	269.21
654.01	645.78	643.10	666.66	1254.90	647.88	646.31	644.84	643.45
46.13	45.02	43.26	50.93	179.37	45.95	45.68	44.93	43.66
495.62	488.49	487.40	507.93	1033.72	489.58	487.04	487.86	485.94
655.06	647.19	645.44	665.80	1191.27	647.75	646.01	647.06	643.86
694.87	690.01	688.82	703.27	1048.98	691.23	689.78	690.58	687.68
629.11	624.86	624.21	635.61	905.66	625.37	624.00	625.95	622.87
760.48	754.90	753.93	768.90	1147.50	755.58	754.09	755.31	753.13

4c. Doubly masked data Ax*B_1 for the first observation.

Time	Censoring	ADH	Age	CD4	Log10PVL	Male	STI	DHU
274.78	270.64	269.05	281.89	260.00	201.97	241.02	474.73	472.13
654.01	645.78	643.10	666.66	623.01	504.49	585.38	1065.30	1059.20
46.13	45.02	43.26	50.93	40.84	13.99	31.00	137.67	136.08
495.62	488.49	487.40	507.93	466.66	361.58	433.68	862.93	859.30
655.06	647.19	645.44	665.80	625.25	521.06	591.57	1021.02	1017.06
694.87	690.01	688.82	703.27	676.52	607.44	653.32	936.66	934.32
629.11	624.86	624.21	635.61	613.64	560.64	595.77	817.39	816.42
760.48	754.90	753.93	768.90	739.15	664.39	715.37	1024.58	1022.12

*Variable abbreviations are the same as those in Table **1**.

effect of age on the adherence to antiretroviral therapy, Table **5a** shows that the masked data and the original data yield exactly the same adjusted odds ratio, Wald confidence interval and p-value corresponding to the age effect. Similarly, in the Cox proportional hazard model to study the effect of age on time to viral load suppression, the masked data also yield the same results for the adjusted hazard ratio on the age effect (see Table 5b). It is worthy pointing out that, because of the attribute transformation, the effects corresponding to the covariates are no longer the same.

CONCLUSION

In this paper, we use a privacy-preserving data collection method to help resolve the privacy protection issue in HIV research. It is shown that the individual data can be completely hidden while the same inference results can still be obtained from the masked data, with the use of common statistical analysis methods. Specifically, orthogonally transformed data enable us to obtain the same results with the use of general linear model and contingency table

Table 5. Results of regression analyses based on the original data and the masked data.

5a. Adjusted odds ratios (AOR) highlighting the effect of age (per 10 years younger) on adherence to antiretroviral therapy in the logistic regression model.

Variable	Results Based on Original Data X		Results Based on Masked Data XB_1B_2	
	AOR (95% CI)	p-Value	AOR (95% CI)	p-Value
Effect of Age	**0.94 (0.28, 3.11)**	**0.91**	**0.94 (0.28, 3.11)**	**0.91**
CD4	0.998 (0.993, 1)	0.43	2.05 (0.36, 11.71)	0.42
Log10PVL	0.53 (0.12, 2.35)	0.41	2.03 (0.37, 11.08)	0.41
Male	1.14 (0.18, 7.41)	0.89	3.62 (0.71, 18.55)	0.12
STI	7.05 (0.52, 95.93)	0.14	0.33 (0.06, 1.86)	0.21
DHU	0.53 (0.07, 3.94)	0.53	0.45 (0.05, 4.11)	0.48

5b. Adjusted hazard ratios (AHR) highlighting the effect of age (per 10 years younger) on time to viral load suppression in the Cox proportional hazards regression model.

Variable	Results Based on Original Data X		Results Based on Masked Data XB_1B_2	
	AHR (95% CI)	p-Value	AHR (95% CI)	p-Value
Effect of Age	**1.30 (0.64,2.64)**	**0.47**	**1.30 (0.64, 2.64)**	**0.47**
CD4	1.00 (0.997, 1.002)	0.72	1.32 (0.58, 3.00)	0.51
Log10PVL	2.89 (1.27, 6.60)	0.01	0.97 (0.41, 2.26)	0.94
Male	0.67 (0.2, 2.22)	0.51	0.49 (0.21, 1.14)	0.10
STI	0.58 (0.18, 1.89)	0.37	3.84 (0.81, 18.25)	0.09
DHU	2.46 (0.55, 11.06)	0.24	1.15 (0.38, 3.49)	0.81

analysis; and certain attribute transformed data enable us to achieve the same statistical inference on the parameter of interest in logistic regression or Cox proportional hazard regression.

It is worthy to note that the new method hides sensitive data with no efficiency loss for statistical inference of binary and normal data, which improves over Warner's randomized response technique, which randomly flips an interviewee's true binary response with a predetermined probability [29-32]. In addition, the new method builds data masking into the data collection device/system so that no additional random device is needed.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Wu SS, Chen S, Burr D, Zhang L. A new data collection technique for preserving privacy. Journal of Privacy and Confidentiality 2014. (In press)

[2] CDC HIV/AIDS Statistics Overview: http://www.cdc.gov/hiv/statistics/basics/.

[3] Boulle A, Schomaker M, May MT, *et al*. Mortality in patients with HIV-1 infection starting antiretroviral therapy in South Africa, Europe, or North America: A collaborative analysis of prospective studies. PLOS Med 2014; 11(9): e1001718.

[4] The Henry J. Kaiser Family Foundation: U.S. Federal Funding for HIV/AIDS: The President's FY 2015 Budget Request.

[5] 2010 Consumer Partnership for eHealth: Protecting Sensitive Health Information in the Context of Health Information Technology.

[6] Rajasingham R, Mimiaga MJ, White JM, Pinkston MM, Baden RP, Mitty JA. A Systematic Review of Behavioral and Treatment Outcome Studies Among HIV-Infected Men Who Have Sex with Men Who Abuse Crystal Methamphetamine. AIDS PATIENT CARE and STDs 2012; 26: 36-52.

[7] Knippels HM, Goodkin K, Weiss JJ, Wikie FL, Antoni MH. The importance of cognitive self-report in early HIV-1 infection: validation of a cognitive functional status subscale. AIDS 2002; 16(4): 681.

[8] Rindskopf DM, Strauss SM, Falkin GP, Deren S. Assessing the Consequences of Using Self-report Data to Determine the Correlates of HIV Status: Conditional and Marginal Approaches. Multivariate Behavioral Research 2003; V38 n3: 325-52.

[9] Kinsey AC, Pomeroy WB, Martin CE. Sexual behavior in the human male. Philadelphia: W. B. Saunders 1948.

[10] Kinsey AC, Pomeroy WB, Martin CE, Gebhard PH. Sexual behavior in the human female. Philadelphia: W. B. Saunders 1953.

[11] Hoff CC, Chakravarty D, Beougher SC, Neilands TB, Darbes LA. Relationship Characteristics Associated with Sexual Risk Behavior Among MSM in Committed Relationships. AIDS PATIENT CARE and STDs 2012; 26: 738-45.

[12] Hadland SE, Milloy MJ, Kerr T, *et al*. Young age predicts poor antiretroviral adherence and viral load suppression among injection drug users. AIDS Patient Care and STDs 2012; 26: 274-80.

[13] Arya M, Amspoker AB, Lalani N, *et al*. HIV Testing Beliefs in a Predominantly Hispanic Community Health Center During the

Routine HIV Testing Era: Does English Language Ability Matter? AIDS Patient Care and STDs 2013; 27: 38-44.

[14] Elkington KS, Bauermeister JA, Robbins RN, *et al*. Individual and contextual factors of sexual risk behavior in youth perinatally infected with HIV. AIDS Patient Care and STDs 2012; 26: 411-22.

[15] Harrison KM, Kajese T, Hall HI, Song R. Risk Factor Redistribution of the National HIV/AIDS Surveillance Data: An Alternative Approach. Pub Health Rep 2008; 123(5): 618-27.

[16] Xiao Y, Song R, Chen M, Hall HI. Direct and Unbiased Multiple Imputation Methods for Missing Values of Categorical Variables. J Data Sci 2012; 10: 465-81.

[17] Phillips DF. Institutional review boards under stress: Will they explode or change? JAMA 1996; 276: 1623-6.

[18] McWilliams R, Hoover-Fong J, Hamosh A, Beck S, Beaty T, Cutting G. Problematic variation in local institutional review of a multicenter genetic epidemiology study. JAMA 2003; 290: 360-6.

[19] Ness RB. Influence of the HIPPA privacy rule on health research. JAMA 2007; 298: 2164-70.

[20] Kim J. A method for limiting disclosure in microdata based on random noise and transformation. American Statistical Association Proceedings of the Section on Survey Research Methods 1986; 370-4.

[21] Kim J, Winkler W. Masking IRS income data on a merged file between 1990 CPS file and IRS income tax return file. American Statistical Association, Proceedings of the Section of Survey Research Methods 1995; 114-9.

[22] Chawla S, Dwork C, Mcsherry F, Smith A, Wee H. Towards Privacy in Public Databases, Theory of Cryptography Conference (TCC); 2005; Cambridge, MA: Springer-Verlag; 556-77.

[23] Rubin DB. Satisfying confidentiality constraints through the use of synthetic multiply imputed microdata. J Off Stat 1993; 9: 461-8.

[24] Burridge J. Information preserving statistical obfuscation. Stat Comput 2003; 13: 321-7.

[25] Gouweleeuw JM, Kooiman P, Willenborg LCRJ, De Wolf, PP. Post randomization for statistical disclosure control: theory and implementation. J Off Stat 1998; 14: 463-78.

[26] Cox LH, Kelly J, Patil R. Balancing quality and confidentiality for multivariate tabular data. Privacy in Statistical Databases 2004, 3050: 87-98.

[27] Muralidhar K, Sarathy R. Data shuffling: a new masking approach for numerical data. Manage Sci 2006; 52: 658-70.

[28] Liu K, Kargupta H, Ryan J. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE T Knowl Data En 2006; 18(1): 92-106.

[29] Ting D, Fienberg SE, Trottini M. Random orthogonal matrix masking methodology for microdata release. Int J Inform Comput Securityroke 2008; 2(1): 86-105.

[30] Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. J Am Stat Assoc 1965; 60, 63-9.

[31] Böckenholt U, van der Heijden PGM. Item randomized-response models for measuring noncompliance: risk-return perceptions, social influences, and self-protective responses. Psychometrika, 2007; 72(2): 245-62.

[32] De Jong MG, Pieters R, Stremersch S. Analysis of sensitive questions across cultures: an application of multigroup item randomized response theory to sexual attitudes and behavior. J Pers Soc Psychol 2012; 3: 543-64.

Received: May 27, 2015 **Revised: September 8, 2015** **Accepted: September 22, 2015**