De-identification of free-text medical records in health information exchange

ZHOU Tian-shu, LI Peng-fei, LI Jing-song*
Healthcare Informatics Engineering Research Center,
Key Laboratory for Biomedical Engineering of Ministry of Education,
Zhejiang University, Hangzhou, 310027, China;
* *To whom correspondence should be addressed; e-mail:* ljs@zju.edu.cn

***Abstract*—Since the health information exchange (HIE) becomes more and more important and numerous systems have been implemented among medical institutions and regions, there also grows the concern of data security and privacy protection. In the prior work, we have designed and developed an international clinical data exchange system named Global Dolphin, technically achieving the goal of health information exchange between China and Japan. However, to put the system into practical use, we have to take the different privacy protection rules into account. In the clinical data exchange implementation, we tried to conform to the stricter privacy protection standards and designed a protected health information (PHI) de-identification system to keep the personal information secure from third parties. Since China and Japan do not yet have detailed rules and guidelines such as the safe harbor method in Health Insurance Portability and Accountability Act (HIPAA), thus we have referenced some of the HIPAA rules and guidelines about PHI, and used a dictionary and regular expressions pattern matching de-identification means to protect personal privacy.**

Keywords - de-identification; PHI; HIE; privacy; data security;

I. INTRODUCTION

Health information exchange (HIE) becomes more and more important in medical services within the exchange of information among heterogeneous systems in medical institutions and regions. The sharing of clinical information could not only facilitate clinic visit, lower medical costs, and maximize the use of limited resources, but also assure the consistency and accessibility of clinical data to make a patient centered service, by which doctors could fast read the important clinical information such as medical and allergic history, to assure the safety and continuity of treatment, reduce human errors, and enhance the entire service qualities of medical industries[1-5].

However, with the rapid expansion of HIE practice and system implementation, there also grows the concern of data security and privacy protection. Due to the development stage of HIE, most clinical data exchange at present occurs between organizations within a regional system, with some occurrences at a national level. Earlier, we have designed and developed an international clinical data exchange system named Global Dolphin, technically achieving the goal of health information exchange between China and Japan, which is one of the first trials of cross-border and cross-language clinical data exchange all over the world[6].

China and Japan, However, have different personal information protection laws, and in order to put the Global Dolphin system into practical use, we need to take the different privacy protection rules into account. In the clinical data exchange implementation, we will always try to conform to the stricter privacy protection standards and designed a protected health information (PHI) de-identification system to keep the personal information secure from third parties. Currently, China and Japan do not yet have detailed rules and guidelines about processing PHI such as the safe harbor method in Health Insurance Portability and Accountability Act (HIPAA)[7, 8], thus, we have referenced some of the HIPAA rules and guidelines about PHI, and used a pattern matching and dictionary de-identification means to protect personal privacy.

II. PHI AND SAFE HARBOR METHOD

Protected health information (PHI) is any information, including demographic information, about health status, provision of health care, or payment for health care that can be linked to a specific individual. This is interpreted rather broadly and includes any part of a patient's medical record or payment history. For instance, one could probably infer to identify an individual through his/her Rh- blood type, age, and race type combination with the public information such as phone book or voter registration database. Under the US Health Insurance Portability and Accountability Act, PHI that is linked based on the following table of 18 identifiers must be treated with special care, and the removal of which in health information referred as the safe harbor method[9].

TABLE I. THE 18 PHIS DEFINED IN HIPAA SAFE HARBOR METHOD

PHI Type	***Description***
Names	Person names
Geographical identifiers	All geographical identifiers smaller than a state
Dates	Other than year, all ages over 89
Phone numbers	Phone numbers
Fax numbers	Fax numbers
Email addresses	Email addresses
Social Security numbers	Social Security numbers
Medical record numbers	Medical record numbers
Health insurance numbers	Health plan beneficiary numbers

PHI Type	*Description*
Account numbers	Account numbers
Certificate/license numbers	Certificate/license numbers
Vehicle identifiers	Including serial numbers, license plate numbers
Device identifiers	Device identifiers and serial numbers
URLs	Web Uniform Resource Locators
IP addresses	Internet Protocol (IP) addresses
Biometric identifiers	Including finger, retinal and voice prints
Full face photographic	Full face photographic and any comparable images
Any other unique numbers	Any unique numbers

According to the US HIPAA's Minimum Necessary Standard and the Japan Act on the Protection of Personal Information, if not for the purpose of treatment-related requests or disclosures of health care information, and if not permitted by the individual, the PHI must be de-identified to meet the privacy protection rules.

In the international clinical data exchange, the free-text medical records need to be translated by translators of a third party. Thus, the PHI in the free-text must be de-identified to prevent the leakage of patients' personal information to non-treatment related third party.

There are two ways to process PHI in the free-text, anonymization and de-identification: 1) anonymization is a process in which PHI elements are eliminated or manipulated with the purpose of hindering the possibility of going back to the original data set[10]; 2) de-identification is a process in which PHI is coded and there is still a link to the original, fully identified data set and not anonymized. After the manual translation, we still need to resume the full text to the medical staff for complete information, thus we use a de-identification and re-identification method to process the free-text in the exchanged clinical data.

III. PATTERN MATCHING AND DICTIONARY

We use a dictionary and regular expressions pattern matching method to process 7 types of PHI in free-text as listed below:

1) Names: patients, relatives, employers, and household members;

2) Ages: all ages over 89;

3) Locations: home addresses, zip code, cities;

4) Hospitals or health institutions: medical facilities, laboratories, and care centers;

5) Dates: other than year;

6) Contact information: phone numbers, emails;

7) Identification numbers: insurance numbers, license numbers and identification card numbers.

A. Dictionary Matching Porcess

We constructed a look-up table of 5 types of PHI (including names, locations, hospitals, contacts, and IDs) whose items were derived from preexistent clinical data and structured data from the current exchanging medical documents, Fig. 1 shows the table definition.

The "ItemName" column preserves the PHI items, such as "Alex Chou"; the "ItemMask" column preserves the mask items, such as "Name"; the "ItemDigest" column preserves the unique identity of the PHI item which used for re-identification and resume the original PHI item; the "ItemType" column preserves the type of PHI item; and the "Encode" column preserves the character encoding type, such as "gb18030" for Chinese characters and "shift_jis" for Japanese characters.

ExpItems.Dict					Table	ExpDict
I	PK	Items		DataType	Length	Attr
		Name	ID			Req / Init
	1	Item SN	ItemSerialNO	Integer		O
		Item name	ItemName	String		O
		Item mask	ItemMask	String		O
		Item digest	ItemDigest	String		O
		Item type	ItemType	String		O
		Encode type	Encode	String		O

Figure 1. Table definition of 5 types of PHI

Other than PHI look-up table, we also need to build a common vocabulary to rule out the false negative identified on the non-PHI items, Fig. 2 shows the table definition.

CommWords.Dict					Table	NonPHIDict
I	PK	Items		DataType	Length	Attr
		Name	ID			Req / Init
	1	Word SN	WordSerialNO	Integer		O
		Word	Word	String		O
		Item digest	ItemDigest	String		O
		Shift	Shift	Integer		
		Word type	WordType	String		O
		Encode type	Encode	String		O

Figure 2. Table definition of non-PHI items

The "Shift" column preserves the offset from the PHI in the common phrase. For example, in the phrase "white flower", the word "white" is in the PHI look-up table as a family name, but in the common phrase it means a color, so the "Shift" would be 1 to indicate the offset from the word "white" in phrase "white flower".

The pattern matching work flow is showed in Fig. 3, de-identification module sans the free-text and looks up in the PHI table to identify a PHI candidate, then expands the offset and determines the PHI type, if the PHI is name, location or institution, then looks up in the common vocabulary to see whether it is a PHI item or a common phrase, otherwise, use regular expression to determine if it is a PHI, at last, the PHI item will be replaced with its mask.

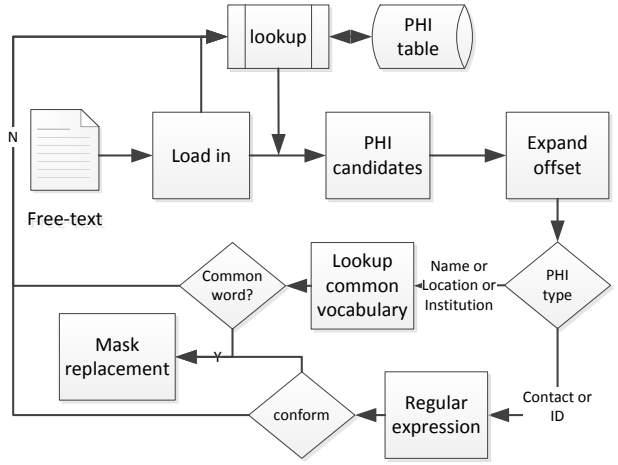

Figure 3. Pattern matching work flow

B. *Regular Expressions Matching Process*

Ages and dates could not use a look-up dictionary pattern matching to identify, so we use a regular expression matching process to locate these two types of PHI. However, these information have multiple expression structures, such as day-month-year or month/day/year, thus we have to build an expression table to traverse all the possible structures, Fig. 4 shows the table definition.

DateFormat.Dict				Table	Date		
I	PK	Items		DataType	Lenght	Attr	
		Name	ID			Req	Init
	1	Date SN	DateSerialNO	Integer		O	
		Date format	DateFormat	String		O	
		Date mask	DateMask	String		O	
		Date digest	DateDigest	String		O	
		Encode type	Encode	String		O	

Figure 4. Table definition of regular expressions

C. *Secure Re-identification Process*

In the re-identification process, we use a secure digital signature method to unique identify the PHI, preventing from the inverse operation of PHI identifier to its original contents, Fig. 5 shows the work flow.

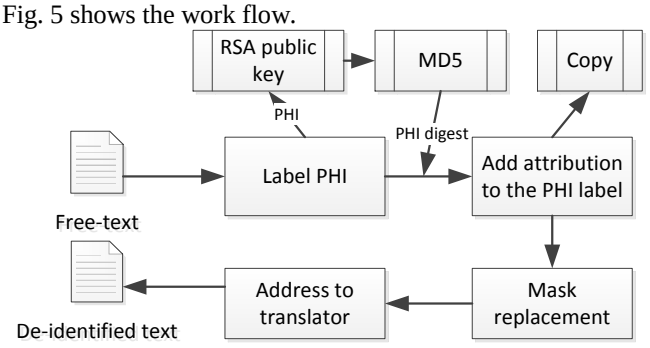

Figure 5. Table definition of regular expressions

IV. RESULTS

As the de-identification system was developed and deployed, we implemented several experiments to test its functionality and identification performance. The data we used in the experiment are shown in table 2, and the identification results are shown in table 3.

TABLE II. EXPERIMENT DATA

PHI types	*quantity*	*distribution rate*
names	207	40.51%
dates	157	30.72%
ages	2	0.39%
locations	78	15.26%
institutions	47	9.20%
contracts	17	3.33%
IDs	3	0.59%

TABLE III. EXPERIMENT REAULTS

PHI types	*locate*	*miss*	*ratio*	*false*	*Precision*
names	215	9	95.65%	17	92.09%
dates	148	11	92.99%	2	98.65%
ages	1	1	50.00%	0	100.00%
locations	68	13	83.33%	3	95.59%
institutions	42	5	89.36%	0	100.00%
contracts	19	0	100.00%	2	89.47%
IDs	3	0	100.00%	0	100.00%

The total PHI identified ratio is 92.37%, and the precision is 95.16%.

V. DISCUSSION AND CONCLUSIONS

As the electronic health record and health information exchange implementation grows all over the world, people concerned more and more about their personal information protection and clinical data safety. If an institution would like to exchange health information with external facilities, it must follow the privacy protection rules according to the local laws; otherwise, it will be illegal to exchange clinical data among different institutions.

As implementing the international clinical data exchange system, and carry out the practical exchange of health information between China and Japan, it's vital to take both countries' privacy laws into account. We designed and developed a de-identification of protected health information in free-text medical record to protect the PHI from leaking to the third-party translator in order to conform to a stricter privacy protection standard and solved the privacy protection differences issue between China and Japan. Since the two countries do not yet have detailed rules and guidelines such as the safe harbor method in HIPAA, we have referenced some of the HIPAA rules and guidelines about PHI, and used

a dictionary and regular expressions pattern matching de-identification means to protect personal privacy.

The experiment shows that the PHI identified ratio is 92.37%, and is higher than one single man's identified ratio which is 81%[11], and the efficiency and cost is far away better than artificial work.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation (Grant No. 61173127) and National High-tech R&D Program (No. 2013AA041201) and Zhejiang University Top Disciplinary Partnership Program (Grant No. 188170*193251101).

REFERENCES

[1] Deas, T.M., Jr. and M.R. Solomon, Health information exchange: foundation for better care. Gastrointest Endosc, 2012. 76(1): p. 163-8.

[2] Kaelber, D.C. and D.W. Bates, Health information exchange and patient safety. J Biomed Inform, 2007. 40(6 Suppl): p. S40-5.

[3] Kuperman, G.J., Health-information exchange: why are we doing it, and what are we doing? J Am Med Inform Assoc, 2011. 18(5): p. 678-82.

[4] Vest, J.R., Health information exchange: national and international approaches. Adv Health Care Manag, 2012. 12: p. 3-24.

[5] Vest, J.R. and L.D. Gamm, Health information exchange: persistent challenges and new strategies. J Am Med Inform Assoc, 2010. 17(3): p. 288-94.

[6] Li, J.S., et al., Design and development of an international clinical data exchange system: the international layer function of the Dolphin Project. J Am Med Inform Assoc, 2011. 18(5): p. 683-9.

[7] Chatterjee, S., Strengthening of the Health Insurance Portability and Accountability Act (HIPAA)-its Role in Public Health. Journal of Theory and Practice of Dental Public Health, 2013. 1(1).

[8] Fielstein, E., S. Brown, and T. Speroff, Algorithmic De-identification of VA Medical Exam Text for HIPAA Privacy Compliance: Preliminary Findings. Medinfo, 2004: p. 1590.

[9] Lafky, D., The Safe Harbor method of de-identification: An empirical test. Fourth National HIPAA Summit West, 2010.

[10] Ohm, P., Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA Law Review, 2010. 57: p. 1701.

[11] Neamatullah, I., Automated de-identification of free-text medical records. BMC Med Inform Decis Mak, 2008. 8: p. 32.