ORIGINAL ARTICLE
Medical Informatics

JKMS

http://dx.doi.org/10.3346/jkms.2015.30.1.7 • J Korean Med Sci 2015; 30: 7-15

A De-identification Method for Bilingual Clinical Texts of Various Note Types

Soo-Yong Shin,[1,2,*] Yu Rang Park,[2,*] Yongdon Shin,[2] Hyo Joung Choi,[2] Jihyun Park,[2] Yongman Lyu,[2] Moo-Song Lee,[3] Chang-Min Choi,[2,4,5] Woo-Sung Kim,[1,4] and Jae Ho Lee[1,2,6,7]

[1]Department of Biomedical Informatics, [2]Office of Clinical Research Information, Asan Medical Center, Seoul; [3]Department of Clinical Epidemiology and Biostatistics, Departments of [4]Pulmonary and Critical Care Medicine, [5]Oncology, [6]Emergency Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea; [7]Division of General Internal Medicine, Brigham and Women's Hospital, Boston, MA, USA

*Soo-Yong Shin and Yu Rang Park contributed equally to this work.

Received: 26 February 2014
Accepted: 29 August 2014

Address for Correspondence:
Jae Ho Lee, MD
Department of Biomedical Informatics, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 138-736, Korea
Tel: +82.2-3010-5875, Fax: +82.2-3010-8126
E-mail: rufiji@gmail.com

Funding: This study was supported by a grant (2013-7205) from the Asan Institute for Life Sciences, Seoul, Korea.

De-identification of personal health information is essential in order not to require written patient informed consent. Previous de-identification methods were proposed using natural language processing technology in order to remove the identifiers in clinical narrative text, although these methods only focused on narrative text written in English. In this study, we propose a regular expression-based de-identification method used to address bilingual clinical records written in Korean and English. To develop and validate regular expression rules, we obtained training and validation datasets composed of 6,039 clinical notes of 20 types and 5,000 notes of 33 types, respectively. Fifteen regular expression rules were constructed using the development dataset and those rules achieved 99.87% precision and 96.25% recall for the validation dataset. Our de-identification method successfully removed the identifiers in diverse types of bilingual clinical narrative texts. This method will thus assist physicians to more easily perform retrospective research.

Keywords: De-identification; Anonymization; Clinical Text; Bilingual Text; Patient Privacy; Medical Informatics; Text Mining

INTRODUCTION

Electronic Health Record (EHR) systems have been widely adopted in the United States (1) and in Korea (2-4). Therefore, research utilizing data obtained from EHR systems has increased due to the ease of accessing a large amount of clinical data (5-8). For example, in our hospital, there were 1,746 requests to extract EHR data for research purposes in 2012 (9). A survey also determined that 64% of clinicians (92 out of 143) have used EHR data for clinical research (10). However, at the same time, patient privacy concerns have arisen. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) defined secondary usage guidelines for medical records, and the Office for Civil Rights recently published guidelines for the de-identification of medical records (11). The Korean government also passed two laws, i.e. the Personal Information Protection Act and the Bioethics and Safety Act, in order to prevent the unauthorized use of medical information; these two laws also proposed the de-identification of personal health information as an alternative to obtaining informed consent from each study participant.

De-identification is an effective method for protecting patient privacy and complying with governmental regulations while improving the convenience of performing research (12). Similar to other medical centers (13), we have been developing a biomedical research platform that uses de-identification to protect patient privacy (14, 15). Diverse automatic de-identification methods have also been proposed to remove the identifiers in clinical notes that are written in free text form (16-22). Because natural language processing (NLP) methods have been developed to manage clinical text and achieved reliable performance (23), most de-identification methods use NLP technology.

However, sophisticated NLP-based methods had not yet been prepared for bilingual clinical text written in Korean and English. First, physicians in Korea write clinical notes in Korean and English. Although there have been several researches regarding non-English text (24-26), most of the previous studies focused on English sentences using NLP-based de-identification meth-

pISSN 1011-8934
eISSN 1598-6357

ods. In particular, there are only a few studies regarding de-identification methods for clinical data in Korea (27-29). Second, most of the text in clinical notes are not full sentences, but rather phrases, and in some instances, these are not grammatically correct phrases. Grammatical NLP methods might suffer from insufficient parsing information. Third, it is difficult to find reliable open-source NLP tools for the Korean language. In our experience, this was the practical limitation of applying sophisticated NLP methods. Regular expression, which is a sequence of characters that forms a search pattern mainly for use in string matching (30), has been employed in our proposed method, since it easily incorporates prior knowledge and demonstrates reliable performance on clinical text processing (31). In addition, its advantages include the speed and ease of use as regular expression syntax is mostly standard across all implementations and regular expressions can usually be transferred to any of other programs with minimal modifications (30).

In this study, we propose a new de-identification method which utilizes regular expression rules to remove the identifiers in bilingual clinical notes written in Korean and English. To cover as many cases as possible, we developed the rules using 6,039 clinical notes of 20 types and validated the rules using 5,000 clinical notes of 33 types. Until now, these two datasets are the most comprehensive datasets for de-identification research.

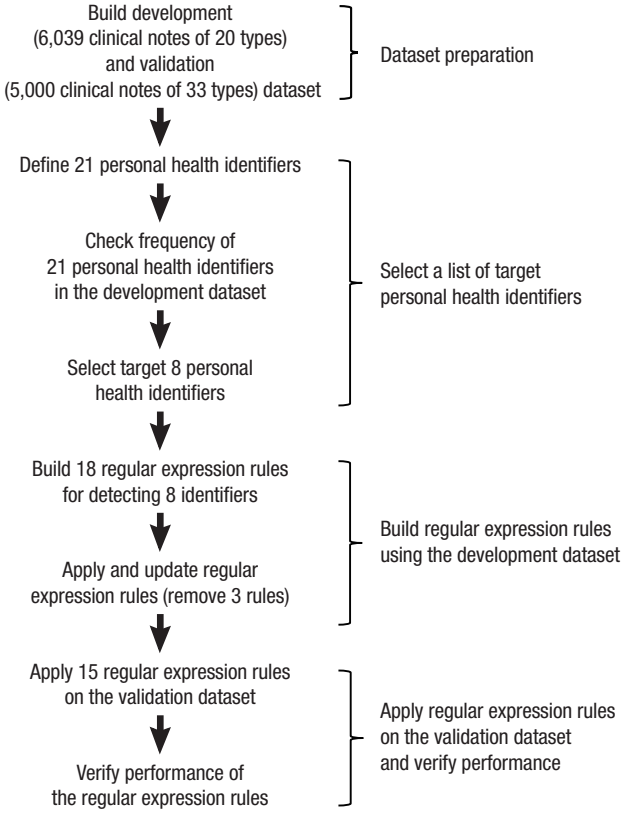

Fig. 1. The process of developing the de-identification method.

MATERIALS AND METHODS

The overall procedure is described in Fig. 1. The following subsections will describe each method in detail.

Dataset preparation

We carefully designed two gold-standard datasets. The first was development dataset and the second was validation datasets, which are composed of clinical notes, in order to develop the regular expression rules and to verify the performance of the rules.

For the development dataset, we manually selected the types of clinical notes. First, we selected patients seen between 2006 and 2011 because our EHR system has been used since 2006. Among these patients, we chose those who revisited the same physicians at our hospital in 2012 as outpatients. This criterion forced the inclusion of more than two clinical notes written by the same physician for each patient. Next, in order to increase the diversity of clinical notes, we chose patients who had admission notes or emergency room notes. As a result, 498 patients with 6,502 clinical notes were selected. Because some clinical notes did not have sufficient extractable free text, we chose 6,039 clinical notes that consisted of 20 different types, including eleven inpatient, three outpatient, and six emergency room notes (Fig. 2). These clinical notes were written by 493 different physicians.

For the validation dataset, we sampled 5,000 clinical notes by stratified random sampling from 14,328,473 clinical notes that included 1,136,005 emergency room notes, 7,933,291 inpatient notes, and 5,259,177 outpatient notes. Based on these proportions, we chose 400 emergency room notes, 2,750 outpatient notes, and 1,850 inpatient notes. For each category, 15 types of inpatient clinical notes, 11 types of outpatient clinical notes, and seven types of emergency room notes were chosen. We sampled clinical notes that had been used more than 3,000 times in order to choose frequently used ones. For each type of clinical note, we chose lengthy clinical notes.

The personal health identifiers in the development and the validation datasets were manually annotated by five annotators consisting of two programmers, one registered nurse, and two medical records administrator. Four annotators including two programmers, one registered nurse, and one medical records administrator, each separately reviewed a quarter of the datasets. After which the other medical records administrator manually annotated datasets again. We also measured discrepant results between the different annotation results between the first annotation (the sum of four annotators) and the second annotation. We calculated the Inter-Annotator Agreement (IAA) score for the development dataset. The IAA score for the development dataset was 0.963 using Cohen's kappa, and which seems to be reliable for annotation results.

http://dx.doi.org/10.3346/jkms.2015.30.1.7

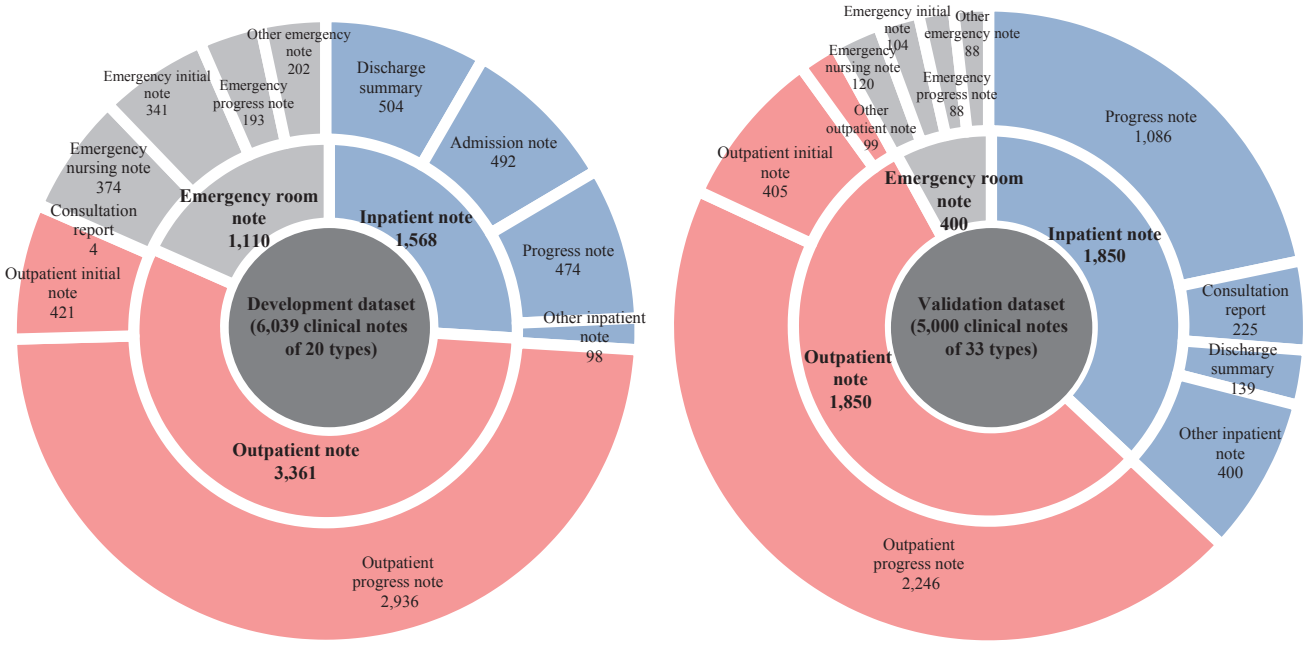

Fig. 2. Development and validation dataset.

Table 1. Institutional Personal Health Identifiers which is modified from Table 2 in Reference No. 14. Adapted after permission

No.	Identifier	Remarks
1	Names	Excludes physician's name Includes information regarding friends and relatives
2	Addresses	Smaller than the sub-municipal level divisions (Dong, -Eup, and -Myeon)
3	Phone numbers	Includes mobile phone and fax numbers
4	Email addresses	
5	Korean resident registration numbers	
6	Foreigner registration numbers	
7	Passport numbers	
8	Health insurance policy numbers	
9	Bank account numbers	
10	Credit card numbers	
11	Certificate/license numbers	Driver's license
12	Vehicle license plate numbers	
13	Patient IDs	Medical record numbers
14	Hospital membership IDs	Hospital homepage, referral system
15	Hospital employee numbers	
16	IP addresses	
17	URLs	
18	Biometric identifiers	Fingerprints, retina, vein, voice prints, and personally identifiable genetic information
19	Full face photographic images and any comparable images	
20	Birth-dates (allowing year and month)	July, 1960 can be used, but July 4, 1960 should be used as July **, 1960
21	Other unique identifying numbers	Pathology numbers

Institutional definition of personal health identifiers

Defined personal health identifiers are required to de-identify narrative clinical text. However, the Korean government has not defined these identifiers in as much detail as HIPAA did (11). Our hospital defined 21 personal health identifiers (Table 1) based on 18 HIPAA identifiers, two Korean regulations, and the ISO standard (ISO/TS 25237:2008) (32).

The information regarding friends and relatives of hospital employees (categorized into "Names" identifiers as in Table 1) was included to increase the patient privacy as even with a masked patient name, a patient could be identified according to that information, e.g. the father of Dr. Kim and the friend of Mr. Park. However, we did not include dates directly related to an individual except birth-dates, i.e., ages over 89 yr or hospital visit

date, in the list of identifiers. The limited dataset definition of HIPAA allows date information to be used for research purposes (33). Based on this limited dataset definition, some organizations, such as the University of California, San Francisco, do not de-identify the date information for research datasets (34). However, to increase the protection of patients' privacy, we masked their birth-date. As in Table 1, only the year and month could be used.

Development of regular expression rule

To select target personal health identifiers among 21 identifiers, we manually check the frequency of personal health identifiers occurring in the development dataset. As a result, the following eight personal health identifiers are selected for de-identifying, i.e. names, addresses, phone numbers, email addresses, Korean resident registration numbers, patient IDs, IP addresses, and birth-dates (allowing only the year and month).

For detecting the above eight targeted personal health identifiers in clinical narrative text, we manually constructed regular expression rules using the following steps. First, we built 18 regular expression rules based on prior knowledge of the annotators. Second, we applied these regular expression rules to the development dataset, after which we updated the regular expression rules. As there were only five identifiers, i.e. patient names, addresses, phone numbers, patient IDs, and birth-dates, in the development datasets, we removed three regular expression rules for detecting three identifiers, i.e. email addresses, Korean resident registration numbers, and IP addresses. Finally, we defined 15 regular expression rules for detecting five personal health identifiers (Fig. 3). The other 16 identifiers were simply removed as they were stored in the structured format of a database system.

Patterns of identifiers	Rules	Examples
Names		
Name of patient	(1) Exact match using list of patients	홍길동님은 ⇒ 이름:****님은
Relatives of patient	(2) Exact match using list of relatives of physicians and employees. If found, remove the physicians or employees name.	
Addresses		
8 provinces + gun (county)	(3) (강원|경기|경상|전라|제주|충청).*군.*[옵면리로].*Wd	현주소: 서울 송파구 ... ⇒ 주소:**********
8 provinces + si (city)	(4) (강원|경기|경상|전라|제주|충청).*시.*[동면리로].*Wd	
6 metropolitan cities & 1 special city	(5) (울산|인천|대구|부산|서울|대전|광주).*구.*[동로길].*Wd	
Phone numbers		
keywords +(: or space) + numbers	(6) (전.*화.*)[0-9]{2,3}[W:WsW;-][0-9]{3,4}[W:WsW;-][0-9]{4}	Tel: 010-1234-5678 ⇒ 전화:***-***-***
	(7) (TEL.*|Tel.*|tel.*)[0-9]{2,3}[W:WsW;-][0-9]{3,4} [W:WsW;-][0-9]{4}	
	(8) (HP|Hp|hp|H.P|H.p|h.p).*[0-9]{2,3}[W:WsW;-][0-9]{3,4} [W:WsW;-][0-9]{4}	
	(9) .*폰.*[0-9]{2,3}[W:WsW;-][0-9]{3,4}[W:WsW;-][0-9]{4}	
	(10) Wd{2,3}[-.Ws]Wd{3,4}[-.Ws]Wd{4}	
	(11) 팩.*스.*[0-9]{2,3}.*[0-9]{2,4}.*[0-9]	⇒ 팩스:***-***-***
	(12) (FAX|fax|Fax).*[0-9]{2,3}.*[0-9]{2,4}.*[0-9]	
Patient IDs	(13) (등록번호|등록 번호|환자번호|환자 번호|환자 정보| ID|id|IDNO|idno|NO|No|Baby|baby|기증자|수혜자| Donor|donor|Recipient|recepient)WD{0,5}Wd{8}	등록번호: 12345678 ⇒ 등록번호:********
Birth dates	(14) (생.*일|birth|BIRTH|Birth).*Wd{2,4} [;:#%Ws]*Wd{1,2} [;:#%Ws] *Wd{1,2} [;:#%Ws] (15) 출.*생.*Wd{2,4}[;:#%/Ws.]*Wd{1,2}[;:#%/Ws.] *Wd{1,2}[;:#%Ws]	생년월일: 44.5.25 ⇒ 생년월일:44.5.***

Fig. 3. Fifteen regular expression rules for de-identification.

Evaluation criteria

To evaluate the performance of the proposed method, precision, recall, and the F0.5 score were used. Precision was defined as the ratio of correctly masked identifiers to the masked data, recall was defined as the ratio of successfully masked identifiers to that of all identifiers, and the F0.5 score was defined as the $(1+0.5^2) \times \frac{\text{Precision} \times \text{Recall}}{(0.5^2 \times \text{Precision})+\text{Recall}}$. As recall is usually more important than precision for the de-identification, the F0.5 score was calculated.

Ethics statement

The study was approved by the institutional review board of our hospital (IRB No. 2012-0623). Informed consent requirement was waived by the board.

RESULTS

Dataset descriptive statistics

In the development dataset, the annotators discovered 1,862 total identifiers (0.37 identifiers per note, Table 2). Of the 21 institutional personal health identifiers, there were only five identifiers, i.e. patient names, addresses, phone numbers, patient IDs, and birth-dates, in this dataset. The most frequent identifier was the birth-date (940, 50.5%), followed by the phone number (781, 41.9%). The remaining identifiers included 108 patient names, 25 patient IDs, and eight addresses. In the validation dataset, there were 773 identifiers in 5,000 clinical notes (0.15 identifiers per note, Table 2). As in the development dataset, the same five identifiers appeared, and the most frequent identifier was also the birth-date (446, 57.7%), followed by the phone num-

Table 2. The distribution of personal health identifiers in the datasets. The total numbers of identifiers in the clinical notes are shown

	Names	Addresses	Phone numbers	Patient IDs	Birth-dates	Total*
Development dataset	108 (5.8%)	8 (0.4%)	781 (41.9%)	25 (1.3%)	940 (50.5%)	1,862
Inpatient clinical notes	7	1	2	17	16	43 (2.3%)
Outpatient clinical notes	86	7	37	4	914	1,048 (56.3%)
Emergency room clinical notes	15	0	742	4	10	771 (41.4%)
Validation dataset	37 (4.8%)	19 (2.5%)	266 (34.4%)	5 (0.6%)	446 (57.7%)	773
Inpatient clinical notes	5	17	32	3	22	79 (10.2%)
Outpatient clinical notes	28	2	7	0	419	456 (59.0%)
Emergency room clinical notes	4	0	227	2	5	238 (30.8%)

*The sum of the percentages may not be 100% due to their being rounded.

http://dx.doi.org/10.3346/jkms.2015.30.1.7

ber (266, 34.4%). The other identifiers were 37 patient names, 19 addresses, and five patient IDs. As it was designed, the development dataset contained more identifiers than the validation dataset in order to assist in the development of the regular expression rules.

When we investigated each identifier in detail, the same birth-dates, phone numbers, and patient IDs sometimes appeared more than once in a single clinical note. In one case, the same patient ID appeared three times. Most of the phone numbers appeared twice (90%). In some cases, the home and mobile phone numbers were written together, whereas in other cases the same phone number was repeated.

Regular expression based de-identification rules

Fifteen rules that covered these five identifiers in the text were chosen as shown in Fig. 3. The development dataset consisted of 6,039 notes written by 493 different physicians in 60 clinical departments. As our hospital had 1,631 physicians as of July 2013, we reviewed the narrative text written by approximately a quarter of these physicians in order to improve the accuracy of the rules.

Names

We developed two rules to mask patient-name-related identifiers. We used two databases, i.e. the basic patient information database to determine the patient names and the employee family information database to identify the relatives of employees. The first rule masked the patient names by searching the basic patient information database which includes patient names. When implementing the practical de-identification system, we used the metadata for the clinical notes in order to reduce the computational time. A clinical note is composed of metadata and narrative text. The metadata of a clinical note indicate a patient's demographic or identifiable information in our EHR system, such as the patient registration number, health insurance policy numbers, etc.

The second rule masked the information regarding patient friends or relatives. We first attempted to detect friends or relatives using keywords. However, as there are too many diverse expressions in the Korean language that indicate friends or relatives, we decided to use the employee family information database. Our hospital maintains this database so as to exempt medical expense if a relative of an employee visits our hospital. If a patient's name was found in this database, we masked the physician or employee's name in the text to hide whose relative the patient was.

Addresses

We developed three rules to de-identify the patient addresses using the information indicating that Korea is divided into eight provinces, six metropolitan areas, and one capital city. These geographical areas were further subdivided into diverse smaller divisions, i.e. Si (city), Gun (county), and Gu (district). Two rules were developed for all eight provinces as these provinces have two different municipal level divisions such as Si and Gun. One rule was implemented for Si and the other for Gun. The third rule was developed for six metropolitan areas and one capital city (Seoul) as those areas have the same municipal level division, i.e. Gu. All three rules are shown in Fig. 3.

Phone numbers

In Korea, phone numbers have special patterns, such that the area code or mobile phone code (2 or 3 digits, starting with '0') and the subscriber number (7 or 8 digits). A hyphen may be present between the area code and the subscriber number or within the subscriber number (***-**** or ****-****). If these patterns appeared after English or Korean keywords such as Tel, HP, Phone, fax or space, the dedicated rules masked the phone numbers. Due to the complexity of the phone number patterns and the diversity of keywords, we devised seven rules to mask the phone numbers.

Patient identifications

We devised one rule to mask patient identifications (IDs), i.e. medical record numbers. When patient IDs are written in the clinical text, there are always special keywords indicating that those digits are patient IDs. The keywords can be found in rule (13) of Fig. 3. As our hospital medical record numbers have eight digits, this rule simply scanned 8-digit numbers after chosen keywords and then masked them. In actual practice, we also used the clinical note metadata as we did with the first rule for names. We masked the 8-digit numbers that matched a chosen patient's medical records number because the selected clinical note contained metadata that provided the patient's medical records number.

Birth dates

We developed two rules to mask the birth-dates using keywords representing birth-dates in either Korean or English. As indicated in Table 1, the year and month were not masked, although the exact birth-date was masked. An example of masking the birth-dates using rule (14) is shown in Fig. 3. Birth-date information, such as "44.5.25" (year.month.day) is masked as "44.5.**" so as to hide the day-related information.

De-identification results

Using the development dataset to develop regular expression rules, we achieved a 99.1% precision, a 98.7% recall, and a 99.0% F0.5 score. The detailed results are shown in Table 3. Of 1,862 identifiers in 6,039 clinical notes, 1,837 identifiers were accurately masked, 17 non-identifiers were incorrectly masked, and 25 identifiers were not masked. Our method de-identified well

Table 3. The performance of the proposed de-identification method in the development and validation dataset

Dataset type	Personal health identifiers		TP	FP	FN	Precision	Recall	F0.5 Score
Development dataset	Identifiers	Names	83	4	25	95.40	76.85	91.01
		Addresses	8	0	0	100	100	100
		Phone numbers	781	1	0	99.87	100	99.90
		Patient IDs	25	6	0	80.65	100	83.89
		IP addresses	0	5	0	0.00	N/A	N/A
		Birth-dates	940	1	0	99.89	100	99.92
	Note types	Inpatient	40	6	3	86.96	93.02	88.11
		Outpatient	1,030	8	18	99.23	98.28	99.04
		Emergency	767	3	4	99.61	99.48	99.58
	Total		1,837	17	25	99.08	98.65	99.00
Validation dataset	Identifiers	Names	23	0	14	100	62.16	89.15
		Addresses	19	0	0	100	100	100
		Phone numbers	257	0	9	100	96.62	99.30
		Patient IDs	5	0	0	100	100	100
		Birth-dates	440	1	6	99.77	98.65	99.55
	Note types	Inpatient	73	0	6	100	92.41	98.38
		Outpatient	444	1	12	99.78	97.37	99.28
		Emergency	227	0	11	100	95.38	99.04
	Total		744	1	29	99.87	96.25	99.12

TP, true positive; FP, false positive; FN, false negative.

phone numbers, birth-dates, and addresses, however, some patient names were missed (25/108 cases). Our method also worked well with emergency room notes (99.6% precision and 99.5% recall) because it accurately removed phone numbers which constituted more than 96% (742 of 771) of the PHIs in emergency room notes (Table 3). For the inpatient notes, our method missed three names (false negative) and misjudged four IP addresses and two names (false positive). For the outpatient notes, our method missed 18 names and misjudged two names and six patient IDs. For the emergency room notes, our method missed four names and misjudged one IP address, one birth-date and one phone number.

The developed rules were verified using the validation dataset. The validation results are shown in Table 3. The proposed regular expression-based method achieved 99.87% precision, 96.25% recall, and 99.12% F0.5 score. Of the 773 identifiers in 5,000 clinical notes, 744 identifiers were accurately masked, one non-identifier was incorrectly masked, and 29 identifiers remained. The address and patient IDs were correctly de-identified using our methods, and some of names (14/37), phone numbers (9/266) and birth-dates (6/446) were not detected using our methods. And one birth-date in the outpatient notes was misjudged.

DISCUSSION

We developed two gold-standard datasets that included 11,039 clinical notes of 33 different types. To our knowledge, these datasets are the largest and most varied clinical narrative text datasets with real identifiers. The previous systems have been verified using only one or two types of clinical notes, such as pathology reports (36), discharge summaries (17, 38), nursing progress notes (17) or outpatient progress notes (39). One study was evaluated using synthetic identifiers, not real identifiers (37). Very few systems have been evaluated using more than two note types (22, 32).

The regular expression rules that we developed successfully removed identifiers in bilingual clinical narrative texts written in Korean and English. After validation with many clinical documents of various types, the proposed regular expression rule-based system was proven to be a good alternative if the annotators have sufficient prior knowledge and there are no other freely available reliable NLP tools. Grammatical NLP methods are difficult to use with bilingual or multilingual texts as it is difficult to find reliable NLP tools to handle multilingual text. Typically, as English terminology in the clinical domain is common in Korea, clinical narrative text is written in Korean as well as in English. Our proposed method may be a good starting point for use in the countries where physicians use English clinical terminology or their native language. However, other researchers should note that we extensively reviewed more than 11,000 clinical notes of 33 types written by approximately 500 physicians, and we used experienced annotators with prior knowledge in order to develop the regular expression rules. Another advantage of the rule-based system is the ease of managing and updating the rules. The de-identification performance will greatly depend on the development (training) dataset regardless of whether a rule-based method or a machine learning method is utilized. With machine learning methods, we must re-train the system or choose a different algorithm to update or improve the performance. However, adding or modifying regular expression rules might be easier than altering other NLP methods.

Compared with the development dataset, the precision of the validation dataset slightly increased and the recall decreased

http://dx.doi.org/10.3346/jkms.2015.30.1.7

Table 4. The performance comparison of the other methods. This Table has been modified from the results in Reference No. 19 and 22

Methods	Document types	Precision	Recall	Others
Our methods	Various clinical documents	100%	92.97%	
MIST2	Various clinical documents	92.79%	92.81%	
MCRF1	Various clinical documents	95.25%	89.86%	
i2b2 de-id challenge 1	Discharge summaries	> 94%	> 94%	
i2b2 de-id challenge 2	Discharge summaries			> 86% (F1-score)
i2b2 de-id challenge 3	Discharge summaries	> 92%	> 92%	
i2b2 de-id challenge 4	Discharge summaries	> 96%	> 96%	
i2b2 de-id challenge 5	Discharge summaries			> 93% (F1-score)
HMS Scrubber	Pathology reports	43%	98%	
Concept-Match	Pathology reports	Low		
VA system	VA compensation and pension examination		81%	99% (Specificity)
MeDS	HL7 message		99.06%	
HIDE	Pathology reports	98.20%	> 96.3%	
MedLEE	Outpatient follow-up notes	3.2%		
MIT system	Nursing progress notes and discharge summaries	75%		
MEDTAG	Various clinical documents		96.80%	
Scrub	Various clinical documents		99%	
UCML system	Various clinical documents			0.97 (AUC)
Regenstrief Institute system	Pathology reports		92.70%	
State De-id	Discharge summaries	99%	97%	

MIST2, MITRE Identification Scrubber Toolkit; MCRF1, Mallet Conditional Random Field; i2b2, Informatics for Integrating Biology & the Bedside; HMS, Harvard Medical School; VA, Veterans Affairs; MeDS, Medical De-identification System; HIDE, Health Information DE-identification; MedLEE, Medical Language Extraction and Encoding; MIT, Massachusetts Institute of Technology; MEDTAG, Medical Document Tag; UCML, University Council of Modern Languages; AUC, Area Under Curve.

due to missed patient names. However, our method exhibited reliable performance compared to the previous results as in Table 4, even though our method handled diverse types of clinical notes and bilingual texts. Table 4 shows the types of the clinical notes as well as the performance metric. The most recent research reported a 95.08% precision and a 91.92% recall for 3,503 clinical notes of 22 types (22). Similar work using a regular expression-based method exhibited a 74.9% precision and a 96.7% recall for the development dataset; the test dataset had an estimated recall of 94.3% even though this method was developed using nursing notes, discharge summaries, and X-ray reports (17). The multilingual system developed by Ruch et al. achieved a 99.4% precision and a 98.5% recall for 800 discharge summary documents (26). However, this method was validated using a small set of documents of a single type.

Our regular expression based de-identification method missed some of the personal health identifiers, in particular the patient name. All of the false negative cases of development data for patient names pertained to information regarding the friends or relatives of hospital staff, i.e. the friend of Prof. Kim and the father of Dr. Kim. As explained in the previous section, we decided to use the employee family information database rather than developing regular expression rules. We, therefore, missed other relatives who were not registered in this database, and friends of employees. Similar to the development dataset, all of the false negative cases of the validation dataset for patient names were with regard to information about the friends or relatives of hospital staff. All of the false negative cases for phone numbers were the phone numbers without the area codes. All of the false negative cases for birth-dates had different formats which were not included in our rules.

In the United States, the discharge summary contains more identifiers than any other type of clinical note (22, 36), however, Table 1 presented something different. In our context, outpatient progress notes, followed by emergency room nursing notes, contained more identifiers than the other note types. Birth-dates were especially dominant in the outpatient progress notes, and phone numbers were dominant in the emergency room nursing notes. Inpatient notes, including the discharge summary, had fewer identifiers than the outpatient and emergency room notes. Interestingly, both the frequency of identifiers and the types of identifier-rich clinical notes differ between those seen in the United States and in our study. The different healthcare systems and different cultures may thus affect the style of clinical narrative text. For example, patient names and visit dates are automatically displayed in our EHR system, physicians do not need to record this information in the narrative text unless they have a special reason for doing so.

The following are potential explanations for the superior performance of our method. First, physicians share similar styles of the clinical narrative text. Only 15 rules can cover the clinical note from approximately 500 physicians. Second, the same identifiers were repeated many times due to the copy-and-paste function. Third, in our cases, the free texts included keywords that were suitable for regular expressions as illustrated in Fig. 3.

Based on the reliable performance of this regular expression-

based method, it was applied to the clinical research data warehouse system which can search, review, and extracts the necessary clinical data in our hospital.

DISCLOSURE

The authors have no conflicts of interest to disclose.

AUTHOR CONTRIBUTIONS

Manuscript preparation: all authors. Manuscript approval: all authors. Supervision of the study: JH Lee. Study design: SY Shin, YR Park. Data preparation: Y Lyu, Y Shin. Manual annotation: Y Shin, HJ Choi, J Park, Y Lyu. Regular expression development: Y Lyu. Personal Health Identifier Definition: SY Shin, HJ Choi, J Park, Y Lyu, MS Lee, CM Choi, WS Kim, JH Lee. Data analysis: SY Shin, YR Park. Data visualization: YR Park.

ORCID

Soo-Yong Shin *http://orcid.org/0000-0002-2410-6120*
Yu Rang Park *http://orcid.org/0000-0002-4210-2094*
Yongdon Shin *http://orcid.org/0000-0002-1144-9458*
Hyo Joung Choi *http://orcid.org/0000-0003-4715-2901*
Jihyun Park *http://orcid.org/0000-0002-1872-8708*
Yongman Lyu *http://orcid.org/0000-0002-9363-252X*
Moo-Song Lee *http://orcid.org/0000-0003-1085-9073*
Chang-Min Choi *http://orcid.org/0000-0002-2881-4669*
Woo-Sung Kim *http://orcid.org/0000-0002-1254-1264*
Jae Ho Lee *http://orcid.org/0000-0003-2619-1231*

REFERENCES

1. The Office of the National Coordinator for Health Information. *Update on the adoption of health information technology and related efforts to facilitate the electronic use and exchange of health information. Available at http://www.healthit.gov/sites/default/files/rtc_adoption_of_healthit_and_relatedefforts.pdf [accessed on 19 February 2014].*
2. Yoon D, Chang BC, Kang SW, Bae H, Park RW. *Adoption of electronic health records in Korean tertiary teaching and general hospitals. Int J Med Inform 2012; 81: 196-203.*
3. Fuad A, Hsu CY. *High rate EHR adoption in Korea and health IT rise in Asia. Int J Med Inform 2012; 81: 649-50.*
4. Ryu HJ, Kim WS, Lee JH, Min SW, Kim SJ, Lee YS, Lee YH, Nam SW, Eo GS, Seo SG, et al. . *Asan medical information system for healthcare quality improvement. Healthc Inform Res 2010; 16: 191-7.*
5. Embi PJ, Kaufman SE, Payne PR. *Biomedical informatics and outcomes research: enabling knowledge-driven health care. Circulation 2009; 120: 2393-9.*
6. Jensen PB, Jensen LJ, Brunak S. *Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012; 13: 395-405.*
7. Yoo S, Kim S, Lee KH, Jeong CW, Youn SW, Park KU, Moon SY, Hwang H. *Electronically implemented clinical indicators based on a data warehouse in a tertiary hospital: its clinical benefit and effectiveness. Int J Med Inform 2014; 83: 507-16.*
8. Shin SY, Kim WS, Lee JH. *Characteristics desired in clinical data warehouse for biomedical research. Healthc Inform Res 2014; 20: 109-16.*
9. Lyu Y, Shin Y, Choi HJ, Park J, Lee MS, Kim HJ, Shin SY, Lee JH. *The analyzing of clinical information requesting pattern for clinical research data warehouse in Asan Medical Center. Proceedings of the Korean Society of Medical Informatics 2013 Spring Symposium2013, p.142-3.*
10. Choi HJ, Ryu HJ, Lyu Y, Shin Y, Park J, Shin SY, Lee JH. *A survey on clinical research using EMR. Proceedings of the Korean Society of Medical Informatics 2012 Spring Symposium.*
11. The Office for Civil Rights. *Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule Available at http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf [accessed on 7 February 2014].*
12. McGraw D. *Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. J Am Med Inform Assoc 2013; 20: 29-34.*
13. Liu J, Erdal S, Silvey SA, Ding J, Riedel JD, Marsh CB, Kamal J. *Toward a fully de-identified biomedical information warehouse. AMIA Annu Symp Proc 2009; 2009: 370-4.*
14. Shin SY, Lyu Y, Shin Y, Choi HJ, Park J, Kim WS, Lee JH. *Lessons learned from development of de-identification system for biomedical research in a Korean Tertiary Hospital. Healthc Inform Res 2013; 19: 102-9.*
15. Shin SY, Lyu Y, Shin Y, Choi HJ, Park J, Kim WS, JH. *L. De-identification method for bilingual EMR free texts The American Medical Informatics Association 2013 Symposium2013, p.1290.*
16. Uzuner O, Luo Y, Szolovits P. *Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007; 14: 550-63.*
17. Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD. *Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008; 8: 32.*
18. Loukides G, Gkoulalas-Divanis A, Malin B. *Anonymization of electronic medical records for validating genome-wide association studies. Proc Natl Acad Sci U S A 2010; 107: 7898-903.*
19. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. *Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010; 10: 70.*
20. El Emam K. *Methods for the de-identification of electronic health records for genomic research. Genome Med 2011; 3: 25.*
21. El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Rose S, Howard J, Gluck J. *De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. J Med Internet Res 2012; 14: e33.*
22. Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, Marsolo K, Jegga A, Kaiser M, Stoutenborough L, et al. *Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2013; 20: 84-94.*
23. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. *Extracting information from textual documents in the electronic health record: a re-*

 http://dx.doi.org/10.3346/jkms.2015.30.1.7

view of recent research. Yearb Med Inform 2008: 128-44.

24. Grouin C, Rosier A, Dameron O, Zweigenbaum P. *Testing tactics to localize de-identification. Stud Health Technol Inform 2009; 150: 735-9.*
25. Velupillai S, Dalianis H, Hassel M, Nilsson GH. *Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. Int J Med Inform 2009; 78: e19-26.*
26. Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. *Medical document anonymization with a semantic lexicon. Proc AMIA Symp 2000: 729-33.*
27. Kim I, Lee J, Kim I, Kwak Y. *A new method of registering the XML-based clinical document architecture supporting pseudonymization in clinical document registry framework. J Korean Institute Inform Sci and Engineers: Software and Applications 2007; 34: 918-28.*
28. Lee HJ, Du R. *Anonymity of medical brain images. Inst Electron Eng Korea 2012; 49: 81-7.*
29. Kwon YJ, Yeon JH, Lee SG. *Anonymization techniques suitable for real medical datasets. Proceedings of Korea Computer Congress 2011, p.80-3.*
30. Jurafsky D, Martin JH. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J.: Prentice Hall, 2000.*
31. Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. *Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. J Am Med Inform Assoc 2006; 13: 691-5.*
32. International Standards Organization. *ISO/TS 25237:2008: Health informatics -- Pseudonymization. Available at http://www.iso.org/iso/catalogue_detail?csnumber=42807 [accessed on 7 February 2014].*
33. National Institutes of health (U.S.) *Department of Health and Human Services. How Can Covered Entities Use and Disclose Protected Health Information for Research and Comply with the Privacy Rule? Available at http://privacyruleandresearch.nih.gov/pr_08.asp [accessed on 7 February 2014].*
34. University of California San Francisco, The Committee on Human Research. *The human research protection program. Available at http://www.research.ucsf.edu/chr/HIPAA/chrHIPAAfaq.asp [accessed on 7 February 2014].*
35. Aberdeen J, Bayer S, Yeniterzi R, Wellner B, Clark C, Hanauer D, Malin B, Hirschman L. *The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int J Med Inform 2010; 79: 849-59.*
36. Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. *Development and evaluation of an open source software tool for de-identification of pathology reports. BMC Med Inform Decis Mak 2006; 6: 12.*
37. Szarvas G, Farkas R, Busa-Fekete R. *State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc 2007; 14: 574-80.*
38. Uzuner O, Sibanda TC, Luo Y, Szolovits P. *A de-identifier for medical discharge summaries. Artif Intell Med 2008; 42: 13-35.*
39. Morrison FP, Li L, Lai AM, Hripcsak G. *Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? J Am Med Inform Assoc 2009; 16: 37-9.*