Chapter 27
Challenges in Synthesizing Surrogate PHI in Narrative EMRs

Amber Stubbs, Özlem Uzuner, Christopher Kotfila, Ira Goldstein, and Peter Szolovits

Abstract Preparing narrative medical records for use outside of their originating institutions requires that protected health information (PHI) be removed from the records. If researchers intend to use these records for natural language processing, then preparing the medical documents requires two steps: (1) identifying the PHI and (2) replacing the PHI with realistic surrogates. In this chapter we discuss the challenges associated with generating these realistic surrogates and describe the algorithms we used to prepare the 2014 i2b2/UTHealth shared task corpus for distribution and use in a natural language processing task focused on de-identification.

27.1 Introduction

Before researchers can use electronic medical records (EMRs) outside of the institution that generated the records, it is critical they remove all protected health information (PHI) from the documents. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) [12] provides "safe harbor" guidelines which define what information is considered PHI. The list of PHI includes information such as patient names, ID numbers, phone numbers, email addresses, and

A. Stubbs (✉)
School of Library and Information Science, Simmons College, Boston, MA, USA
e-mail: stubbs@simmons.edu

Ö. Uzuner • C. Kotfila
State University of New York, Albany, NY, USA
e-mail: ouzuner@albany.edu; ckotfila@albany.edu

I. Goldstein
Department of Computer Science, Siena College, Loudonville, NY, USA
e-mail: igoldstein@siena.edu

P. Szolovits
Department of Computer Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: psz@mit.edu

A. Gkoulalas-Divanis, G. Loukides (eds.), *Medical Data Privacy Handbook*,
DOI 10.1007/978-3-319-23633-9_27

Admission Date : **06/07/99** Discharge Date : **06/13/1999** HISTORY OF PRESENT ILLNESS : Mr. **John Smith**is a 60 year old male was last admitted in early **January** to **Massachusetts General** with chest pain. He was attended by Dr. **Burke**, then later treated by Dr. **Amy Pagan** and **Luke Strauss**, RN. **Smith** is a **bartender** who primarily works at **Publick House**, but sometimes works weekdays at **Cheers**.	Admission Date : **06/10/72** Discharge Date : **06/16/2072** HISTORY OF PRESENT ILLNESS : Mr. **Thomas Benson**is a 60 year old male was last admitted in early **January** to **Orlando Regional** with chest pain. He was attended by Dr. **Newcomb**, then later treated by Dr. **Yuridia Joy** and **Milo Brock**, RN. **Benson** is a **Food Service Manager** who primarily works at **Ideal Industries**, but sometimes works weekdays at **JITB**.

Fig. 27.1 A fabricated sample medical record before and after surrogate generation

so on. We refer to the process of removing PHI from EMRs as "de-identification". Researchers tasked with de-identifying EMRs face two challenges: (1) Identifying all the PHI in a medical record, and (2) replacing the PHI with some type of placeholder or surrogate.

In charts and other structured medical data, it may be sufficient to replace a patient's name with a generic placeholder such as "[**PATIENT NAME 12345**]". However, in data that is intended to be read naturally, such as hospital discharge summaries and correspondence between doctors, it is important that the de-identification process maintains the discourse structure and readability of the original text. This readability is important both for humans who may want to make use of the text, but also for natural language processing systems that may use the text for training.

We refer to replacing identified PHI with natural-sounding replacement data as "surrogate generation", though it is sometimes referred to as "re-identification"[1] [23]. We maintain that the surrogate generation process should be implemented in such a way that the replacement data retain the same forms as the original and, as much as possible, the same internal temporal and co-reference relationships.

Figure 27.1 shows a sample record that has gone through the surrogate generation process. The left side shows the "original" text (a medical record fabricated for this chapter); the right side shows the processed record.

In this chapter we focus on the task of surrogate generation by discussing related work (Sect. 27.2), the HIPAA definitions of PHI and whether they are sufficient for

[1]Somewhat confusingly, "re-identification" is also sometimes used to refer to determining a person's true identity from de-identified data [7], so we avoid that term for the remainder of this chapter.

true de-identification (Sect. 27.3), the data used for this case study (Sect. 27.4), and the difficulties associated with generating realistic surrogate PHI in clinical data and the approaches we took in generating surrogate PHI for the 2014 i2b2[2]/UTHealth[3] natural language processing (NLP) shared tasks (Sect. 27.5) [25, 30]. Finally, we discuss potential errors introduced by the surrogate generation process (Sect. 27.6) and the relationship between de-identification and surrogate generation (Sect. 27.7).

27.2 Related Work

Studies have shown that the majority of the U.S. population can be identified using only their gender, date of birth, and ZIP code [10, 28]. Another recent study found similar results for Canadian residents using records of people's addresses over an 11 year period [7]. If a patient's true identity were to be reconstructed from poorly de-identified medical records, that breach of trust could potentially endanger future research that relies on sharing de-identified medical records.

In order to test the efficacy of the 18 HIPAA categories for protecting patient identities, Lafky et al. [17] performed a study to determine how easily de-identified records could be linked back to the patient's true identity. They obtained (with IRB approval) a set of approximately 15,000 de-identified patient records. These patients were all from a minority ethnic group, which the researchers thought would make the identification process easier. They then obtained a list of individuals from a commercial data repository; the list was matched to the ethnic group and geographic area (the specifics of the information in this list were not discussed). They identified unique patient records, and then matched them against the commercial data. Once they thought a match was found, another team verified their guess against the original records. In the end, only two patients of the 15,000 were correctly identified. Lafky's research demonstrates that eliminating or pseudonymizing the 18 HIPAA categories of PHI makes the chance of determining patients' true identities very low, although not entirely impossible.

In a more recent study, Meystre et al. [22] explored whether doctors could recognize their own patient's de-identified notes. The authors used their own automated system to remove PHI from recent clinical notes (1–3 months old), then asked physicians to try to identify their patients. None of the doctors correctly identified their patients.

These studies show that good recognition of PHI is critical for patient protection, and indeed, much of the research in the area of EMR de-identification is in building systems that locate and categorize PHI in EMRs. As Li et al. [19] recently noted, Conditional Random Field algorithms (CRFs) [16] are a prominent approach to finding PHI and similar tasks, such as named entity recognition. Gardner and Xiong

[2]Informatics for Integrating Biology and the Bedside.

[3]University of Texas Health Science Center at Houston.

[8] report similar findings, and this observation was reinforced during the 2014 i2b2/UTHealth NLP shared task. This task featured a track on de-identifying clinical narratives [25, 30], and of the top ten systems, five of them used CRFs to identify PHI [24].

However, a CRF trained on gold-standard PHI data is not necessarily enough for complete de-identification, as the results from the 2014 i2b2/UTHealth shared task show [24], and researchers have been looking at other methods to improve automated systems. Interesting recent approaches include augmenting models with data from public medical texts [20] and clustering clinical narratives by complexity prior to de-identification [19].

While PHI identification is a critical component of the surrogate generation process, the focus of this chapter remains on surrogate generation itself. We refer readers interested in de-identification systems to two recent review articles on the subject: Kushida et al. [15] and Meystre et al. [21], as well as Chap. 26 in this volume.

For the remainder of this chapter we focus our attention on surrogate generation systems and procedures that researchers have used to de-identify authentic narrative EMRs to make the records available for researchers outside their home institutions. As we noted earlier, some datasets make use of generic placeholders or other forms of obfuscation when removing PHI [1, 2, 8, 11, 13]. This chapter focuses on surrogate generation techniques that preserve the readability of the natural language found in the original documents, so the remainder of this section predominantly examines surrogate generation systems in that paradigm, and the methods they used.

One of the first surrogate generation systems was Scrub [27], a system that matched the format of the replacement text to the original text. For dates, Scrub would group the detected dates to a point such as the first of the closest month. For names, Scrub used a look-up table, so that the same name in the original file would always be replaced with the same surrogate.

In creating the MIMIC II Clinical database, which is part of PhysioNet [9], Clifford et al. [3] used a two-step approach. First, they used the system created by Neamatullah et al. [23] to identify the PHI. Next, Neamatullah et al. describe generating realistic surrogate PHI to test their de-identification system, but the final version of the full MIMIC II corpus "scrubbed" the PHI by replacing the identified PHI with placeholders in the format "[** (data) **]". "(data)" represents either a date-shifted piece of temporal information, or a marker to indicate what type of PHI it replaced, such as "First name 213" or "Hospital 57". Each set of patient records (both narrative and tabular) has its own randomly-assigned date shift.

Douglass et al. [6] annotated PHI by hand and replaced them with realistic surrogates in a set of 2646 nursing notes from MIMIC II. The authors state that they shifted dates by a random number of weeks and years, while preserving days of the week. They generated names by mixing and matching from a list of Boston residents, and replaced locations with randomly selected small towns or, in the case of hospitals and wards, with information about a fictitious hospital. They maintained co-reference by making sure repeated mentions of the same authentic PHI were

replaced by the same surrogate PHI. At the end of surrogate generation, a human reviewed the suggested surrogates and had the option to modify the surrogate PHI to ensure that it was reasonable.

Uzuner et al. [29] used an earlier version of the realistic surrogate generation methodology we describe in this chapter, and one similar to that used by Douglass et al. [5]. Specifically, for strings such as ID numbers and phone numbers, they replaced the existing digits and letters with randomly-generated ones. For dates they retained the relations between times by offsetting all dates in a record by the same number of days, and ensured that the surrogate dates were properly formatted. For names of people and places, they randomly generated names by selecting syllables from existing names and mixing them together. Finally, to deliberately introduce ambiguity into their surrogates, they replaced some of the randomly-generated person names with medical terms, such as disease names and interventions. We modified the system used by Uzuner et al. [29] to create the system described in this chapter.

Deleger et al. [4] also used realistic surrogates. They generated surrogate names by randomly selecting male, female, unisex, and surnames from lists compiled from the US Census Bureau. They obfuscated phone numbers, ID numbers, and email addresses by randomly selecting new digits or letters as needed. Deleger et al. took a somewhat different approach to generating dates and locations than the other surrogate generation projects we have discussed so far. Rather than date-shifting by some amount of time or randomly selecting locations from a pre-existing list, they used the PHI in the corpus itself to compile lists of the different types of locations (streets, cities, etc) and different date formats ("November 2, 2013", "4/27/03"). Then, they shuffled dates and parts of locations between documents in the corpus.

Removing all PHI from a record and replacing it with surrogate information protects patient privacy, but surrogate PHI is not always perfect. Yeniterzi et al., [31] recently compared the performances of a machine learning-based de-identification tool when trained and tested on different combinations of surrogate and authentic (the authors refer to it as "original") PHI. They found that when they trained the system on surrogate ("resynthesized") PHI, the system did not perform as well on authentic PHI because of the regularization imparted by the surrogate generation process.

In fact, the resynthesized-to-authentic model had the lowest performance of the four comparisons they performed, with f-measures ranging from 0.47 to 0.81. Authentic-to-authentic and resynthesized-to-resynthesized both did well (0.93–1.00 and 0.96–0.99, respectively), with the third-best performance on authentic-to-resynthesized (0.78–0.89). The fact that the systems performed best when trained on similar data suggests that, in order to provide the best out-of-the-box de-identification on authentic records, de-id systems must either be trained on authentic data (which is rarely possible) or that the resynthesized data must mimic natural language as closely as possible.

Using realistic surrogates maintains the discourse structure of the documents and helps train machine learning systems for de-identification of authentic PHI. However, most papers that discuss systems that generate surrogate PHI dedicate

only a paragraph or two to the surrogate generation process. In this chapter, we present a solution that generates realistic surrogate PHI while maintaining both coreference chains and temporal continuity within the narrative, and benefits from human oversight for best results. The remainder of this chapter discusses our surrogate generation method in detail, including the algorithms we used for certain types of PHI, the difficulties posed by the narrative texts, and the need for human intervention in the surrogate generation process. We hope that by creating better surrogate PHI for publicly-available data sets that we can improve the performance of de-identification systems designed on these data even on authentic medical records.

27.3 PHI Categories

HIPAA defines 18 categories of PHI of "the [patients] or of relatives, employers, or household members of the [patients]," that must be removed from medical records.[4] These categories are described in Table 27.1.

As written, categories 1–17 do not include the names of doctors or other medical personnel, names or locations of hospitals or medical facilities, or any other information that identifies a person who is not a patient or directly related to the patient.

However, in our experience with de-identifying medical records, and based on the aforementioned studies on determining the real identity of individuals with minimal information [7, 10, 28], we have found it beneficial to adopt a risk-averse policy towards de-identification and surrogate generation. Our goal is to ensure patients' privacy and to protect their identities, therefore it behooves us to make our best effort to remove any of the information in a record that a malicious person could use to identify a patient.

Consider the situation where an EMR mentions that a patient sees "Dr. Lau of Belfast Hospital", but upon investigation one can learn that not only does Dr. Lau only see a few patients a year, she only treats patients with paranoid schizophrenia, and she was only at Belfast Hospital from 2001–2002. In this case, information about the doctor and the hospital, when triangulated with external knowledge from other sources, could lead to patient identification.

To minimize such risks, we expand our definition of the HIPAA category 18 to include the following:

- All person names in a document, including hospital staff and their user names;
- All locations, including states, countries, geographical areas (e.g., "The Northeast"), landmarks (e.g., "the Grand Canyon"), and non-generic hospital departments;
- Organizations (e.g., Simmons College, Google);
- All portions of dates, including years;

[4] 45 CFR 164.514.

Table 27.1 HIPAA's list of PHI categories

1. Names;
2. All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:
   a. The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and
   b. The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
4. Telephone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. Web Universal Resource Locators (URLs);
15. Internet Protocol (IP) address numbers;
16. Biometric identifiers, including finger and voice prints;
17. Full face photographic images and any comparable images;
18. Any other unique identifying number, characteristic, or code.

- Pager numbers;
- Any of the PHI types included in categories 4–17 that apply to people or organizations other than the patient or the patient's associates;
- Professions held by the patient or people associated with the patient (this does not include the professions of the hospital staff);
- Any other information that potentially indicates a patient's location or a specific time in the document, such as references to historic events.

Figure 27.2 shows a fabricated medical record before surrogate generation; Fig. 27.3 shows the same record after surrogate generation.

We generate surrogates for PHI in all of the HIPAA categories, including category 18. The generated surrogates maintain co-reference between named entities (e.g., people, locations) by replacing authentic PHI that occurs more than once with the same surrogate, and they maintain temporal relationships by shifting all the dates in a patient's records forward by the same interval. We discuss specific details of the surrogate generation process in the following sections, as well as the issues surrounding generating appropriate surrogate information for all the PHI categories, and the effect these surrogates have on certain types of medical research.

```
Record date: <DATE>2011-01-14</DATE>

INFECTIOUS DISEASE ASSOCIATES
<HOSPITAL>CURTIS MEDICAL CENTER</HOSPITAL>

Patient: <PATIENT>Iles, Lois</PATIENT>
Attending: <DOCTOR>Riley, Todd M.</DOCTOR>

<AGE>70</AGE>y/o F was seen in ID clinic on
<DATE>7/20</DATE>, <DATE>9/27</DATE>, <DATE>11/9</DATE>,
and today (<DATE>1/14/11</DATE>).
PMH: DMII (dx late <DATE>90s</DATE>, last A1c = 5.6 in
<DATE>2/10</DATE>)
[...]
Ms. <PATIENT>Iles</PATIENT> was seen and examined with Dr.
<DOCTOR>Tillman</DOCTOR>.
<DOCTOR>Todd Riley</DOCTOR>, MD pager #<PHONE>03268</PHONE>
cc: <DOCTOR>DANIELLE TOMPKINS</DOCTOR>,
<HOSPITAL>CMC</HOSPITAL> Internal Medicine
Signed electronically by
<DOCTOR>Todd Riley</DOCTOR>, MD
<USERNAME>TMR42</USERNAME>
```

Fig. 27.2 Fabricated EMR prior to surrogate generation; PHI are delineated with XML tags

27.4 Data

We developed and tested the surrogate generation algorithms described here over years on two data sets: the 2006 and 2014 i2b2/UTHealth shared task corpora. The 2006 corpus consists of 889 medical discharge summaries [29], which are narrative descriptions of a patient's hospital stay, written at the time of discharge. The corpus contains 19,498 instances of PHI from eight PHI categories: patients, doctors, locations, hospitals, dates, IDs, phone numbers, and ages [29].

The 2014 i2b2/UTHealth shared task corpus is made up of 1304 longitudinal medical records (records that refer to the same patient over a period of time) for 296 patients [14]. In addition to discharge summaries, the 2014 corpus contains other types of narrative medical records, such as inpatient notes and correspondence between specialists and primary care physicians. The 2014 i2b2/UTHealth corpus contains 28,872 instances of PHI [25], and uses the expanded list of PHI categories we described in Sect. 27.3.

The 2006 corpus provided the original data on which the surrogate generation algorithms were tested; we then refined the algorithms for use with the 2014 corpus, as it contained a wider variety of both record types and PHI. The rest of this chapter focuses on how we applied surrogate generation to the 2014 data set; a discussion of the surrogate generation process on the 2006 data can be found in Uzuner et al. [29].

Both the 2006 and 2014 corpus files come from Partners Healthcare, and the Institutional Review Boards (IRBs) of Partners Healthcare, Massachusetts Institute of Technology, and the State University of New York at Albany approved the data

```
Record date: <DATE>2094-02-06</DATE>

INFECTIOUS DISEASE ASSOCIATES
<HOSPITAL>HARRISON COUNTY INFIRMARY</HOSPITAL>

Patient: <PATIENT>Kuhn, Naomi</PATIENT>
Attending: <DOCTOR>Robertson, Carlos U.</DOCTOR>

<AGE>70</AGE>y/o F was seen in ID clinic on
<DATE>8/12</DATE>, <DATE>10/20</DATE>, <DATE>12/2</DATE>,
and today (<DATE>2/06/94</DATE>).
PMH: DMII (dx late <DATE>70s</DATE>, last A1c = 5.6 in
<DATE>3/93</PHI>)
[...]
Ms. <PATIENT>Kuhn</PATIENT> was seen and examined with Dr.
<DOCTOR>Chung</DOCTOR>.
<DOCTOR>Carlos Robertson</DOCTOR>, MD pager #<PHONE>86962</PHONE>
cc: <DOCTOR>FAYE COX</DOCTOR>,
<HOSPITAL>HCI</HOSPITAL> Internal Medicine
Signed electronically by
<DOCTOR>Carlos Robertson</DOCTOR>, MD
<USERNAME>CUR25</USERNAME>
```

Fig. 27.3 Fabricated EMR after surrogate generation; PHI are delineated with XML tags

preparation we describe here, as well as the release of this data under a Data Use Agreement (DUA). The 2014 corpus data will be made available to researchers in November 2015 from http://i2b2.org/NLP; the 2006 data is already available at the same location.

27.5 Strategies and Difficulties in Surrogate PHI Generation

For many of the PHI categories, generating realistic surrogates is relatively simple. For example, phone and fax numbers, URLs, email addresses, and ID or account numbers (i.e., HIPAA categories 4–16), whether they are for patients or medical staff, are usually alphanumeric strings of numbers, letters, and a few special characters such as hyphens, parentheses, periods, and the 'at' character (@). In these cases, it is simple to not only replace the authentic PHI with randomly-generated surrogates, but it is also relatively easy to maintain co-references.

Figure 27.4 presents a summary of the algorithm we used to replace PHI represented by numeric or alphanumeric sequences. By keeping track of the category of each PHI, we were able to maintain co-reference between elements of the same category without revealing similarities between PHI of different categories. For example, given a document with the phone numbers (555) 123–4567 and 123–4567, as well as the medical record ID number 1234567, we can easily maintain co-reference relationships between the two phone numbers, but we would

1. Isolate the string of numbers and digits; check to see if that string or a substring already has a replacement in the appropriate PHI category.
2. Replace any digit by a randomly-generated digit.
3. Replace any lower-case or upper-case alphabetic character by a randomly-generated lower-case or upper-case, respectively, alphabetic character.
4. Leave other characters and spacing alone.

Fig. 27.4 Generic algorithm for replacing alphanumeric strings

randomly replace the medical record ID digits without referring to the already-replaced phone PHI.

Naturally, we implement more specific rules for various categories of PHI. For example, phone numbers have a restriction that they cannot start with 0, and the substring rule does not apply to medical record numbers or other account or ID numbers. This replacement algorithm resulted in unrealistic email addresses and URLs, however. For example: "rgs44@newhospital.org" might be replaced with "bhg10@jrbsotnwwx.hrh". Given the small number of email addresses and URLs in the corpus, we generated surrogates for these by hand, though later implementations of this system may involve more complex solutions to this problem, involving matching pieces of addresses with names and hospitals found in the text.

While the algorithm in Fig. 27.4 is relatively simple, not all co-referent PHI are so easily replaced. The remainder of this section addresses some of the challenges we faced with replacing more complex PHI, as well as the approaches we took to address them. We present these PHI in order of HIPAA categories that most closely match them. Because we are discussing only text-based files, HIPAA categories 16 and 17 (biometrics and facial photographs) are not included in this discussion.

One challenging aspect of the 2014 i2b2/UTHealth NLP shared task corpus that augmented the complexity of generating surrogate PHI is the longitudinal nature of the records in the corpus. We mitigated the problem of cross-document co-reference by merging all of the records for a single patient into a single file for the purposes of surrogate generation. We then processed all of these records together, thereby ensuring that co-references would not be lost between the patient's records. However, having multiple records per patient increased the potential for variations and misspellings of co-referent PHI that our surrogate generator had to handle.

27.5.1 HIPAA Category 1: Names

As previously noted, the PHI in our corpus include the names of medical professionals in addition to patients and their relations. Names provide an interesting challenge in surrogate generation because they can, and do, occur in many different forms within the same document. Also, as previously noted, increasing the number

of documents about a single patient correspondingly increases the number of co-referent PHI. Over the course of multiple documents, a single patient could, for example, be referred to as:

- Vasquez, Angela
- Angela Vasquez
- Angie
- Ms. Vasquez
- Angela M. Vasquez
- and various misspellings of these names, such as "Angel" or "Vazquez".

Nicknames and misspellings are particularly tricky for an automated system to catch. While it is likely that "Angie Vazquez" and "Angela Vasquez" are the same person, it is also possible for these two names to refer to two different people, e.g., a patient and a doctor, something that human readers may be able to distinguish but a simple surrogate generation system may not.

Another layer of challenge is that, in the i2b2 corpora, doctors often initial the bottom of records to indicate they are complete, and sometimes add their hospital system username after their own name. This means that in addition to maintaining co-reference between full names, we needed to be able to identify initials and map them appropriately to the full names, while differentiating them from acronyms that may look exactly the same as these initials.

Finally, we wanted to maintain the genders of the names in the original documents as much as possible. To help with this task, we obtained information from the US Census Bureau and split the information into four dictionaries for lookup: (1) surnames (2) female names (3) male names (4) unisex names. With this resource, we approached surrogate name generation with the algorithm summarized in Fig. 27.5.

Pre-mapping the letters of the alphabet to letters that we then used to select surrogates had many advantages. First, it allowed us to deal easily with ambiguous co-references in the text without having to make leaps of inference. For example, Fig. 27.2 refers to Drs. Tillman and Thompkins. If at some point the narrative also referred simply to "Dr. T.", our system would not have to attempt to disambiguate the reference. Since "Tilman" and "Tompkins" are both replaced with names starting with the same letter ("C" in Fig. 27.3), the ambiguity in the original document is maintained in the surrogate as well. Similarly, the alphabetic mappings allowed us to easily replace initials and usernames without having to infer which doctor might have signed the document.

While overall this algorithm worked fairly well, there were a few problems with certain circumstances. First, we lacked sufficient support for nicknames and misspellings, and so this type of co-reference had to be corrected by hand. Future implementations will utilize a nickname dictionary as well as some rules involving Levenshtein distances [18], which will help identify small errors or misspellings. Second, randomly creating mappings for letters sometimes led to situations where a letter that occurs commonly in names, such as S, would be mapped to a letter that is much less common, such as X, meant that our dictionary would sometimes be

1. Given a record or set of records for a single patient, create two random mappings between letters in the alphabet: one for given names (first, middle) and one for last names (i.e., A mapped to R, B mapped to E, etc.).
2. For each PHI in the name category, normalize the format into categories that correspond with first name, middle name(s), surname, and suffix.
    a. Check to see if any of these PHI have already been mapped to surrogates.
    b. If they have, use those mappings.
    c. If they have not:
        i. Check against the name dictionaries to determine if the name is most likely male, female, unisex, or a surname.
        ii. Use the alphabet mapping for the appropriate name type to randomly select a new surrogate. For example, if "Angela" is determined to be female and A is mapped to R for this document, select a new name from the female dictionary that begins with R.
3. If the name cannot be normalized, assume it is a set of initials and replace them based on the alphabetic mappings. If the name is a username (initials followed by numbers), map the initials and generate random surrogate numbers.
4. Place the new PHI/surrogate mappings in a lookup table.

Fig. 27.5 Algorithm for generating replacement names

forced to re-use the same surrogates for different PHI. This lack of names starting with certain letters potentially added co-references and/or ambiguity where none existed in the original file. One solution to this would be to set the letter to letter mappings to roughly follow distributions of the census data, though this approach could potentially lead to a PHI leak, as it would reveal information about the authentic PHI.

Finally, randomly selecting first, middle, and last names from dictionaries sometimes leads to extremely improbable surrogate names, especially if the system fails to accurately categorize a name as a given name or a surname. One notable example from an early test of our surrogate system led to the creation of a doctor named "Pagan Trout". Unfortunately, Dr. Trout was not included in the final i2b2/UTHealth corpus.

27.5.2 HIPAA Category 2: Locations

The different location types we addressed were: countries, states, ZIP codes, cities, streets, hospitals, organizations, hospital departments, room numbers, and other locations, such as landmarks and geographic regions.

With the exception of ZIP codes, which we changed by using the algorithm in Fig. 27.4 to generate surrogate PHI for locations we created pre-compiled lists of different location types, and randomly selected from those when generating surrogates. However, we had to implement special rules for certain location types. For example, it is common for hospital names to take many different forms in even

a single medical record. “Massachusetts General Hospital” might be referred to by its full name, as “Mass. General”, as “Mass. Gen.” or simply as “MGH”. Therefore, the algorithm for hospital name replacement looked for possible abbreviations and modified the surrogate output to match the format of the original PHI.

We included hospital department names in our list of PHI in order to ensure that a hospital could not be identified by a department name that is unique. We made department names less identifiable by mapping them to a list of generic department names. For example, the department of “anesthesia, critical care, and pain medicine” would become “anesthesiology”. We left alone department names that were already generic (i.e., “Emergency Department”, “Oncology”). Any department that could not be mapped to a more generic version we adjusted by hand.

We also needed to correct demonyms (name for persons from a country or other location) by hand, as the software did not check for them. If a person were described as “from Armenia” in one EMR and “Armenian” in another, we would check to make sure that the surrogate country of origin was the same in both EMRs, and that the demonym took the appropriate form.

27.5.3 HIPAA Category 3: Dates and Ages

One of the tasks in the 2014 i2b2/UTHealth challenge involved a temporal analysis of each patient’s medical records, and so we prioritized maintaining the temporal relationships between all the dates in a patient’s set of records while still obfuscating the authentic dates in the record. We did this by shifting all the dates in a record into the future by the same interval, but randomly selecting the interval for each new patient. Thus we were able to maintain continuity for each patient without revealing any identifiable dates.

In order to shift all the dates in a patient’s records consistently, we first had to identify all of the dates in the records and convert them to a standard format. While co-reference is not a problem with dates in the way that it is with names, converting dates to a standard format still poses a challenge. A date such as “September 29, 2013” is easy to parse and standardize to 2013-09-29, but many other formats are more difficult to interpret. For example, if the same date were represented as 09/29/2013 we could easily tell that the format is mm/dd/yyyy, as “2013” is not a valid month or day, and “29” is not a valid month. However, the date 11/10/13 is ambiguous. While the “13” represents the year in most date formats, whether the “11” represents the 11 month or the 11th day is unclear. For our system, we first introduced logical constraints: a month cannot be a number greater than 12, and a day cannot be greater than 31. In cases where this logic did not help, such as “11/10/13”, we assumed a mm/dd/yy representation, which is more common in America where these records originated. In order to interpret the year, we assumed that any two-digit year of 20 or less applied to the twentyfirst century, while years ending in 21 through 99 applied to the twentieth century.

An ambiguity that we found less easy to resolve was that of two-integer dates, such as "04/03". Even if we assume the American convention that this date represents April 3rd rather than March 4th, we are still left with the possibility that the date could represent April of 2003. Our manual review of these dates in the original texts revealed that both uses of that format (i.e., mm/dd or mm/yy) occur in the patient records, and we could not implement a rule to automatically shift these dates. In the end, our system marked these dates as ambiguous and a human had to interpret and shift them manually.

Other dates, such as named holidays (i.e., Christmas, Halloween) we mapped to calendar dates whenever possible. Any date that could not be converted into a standard format we marked as "unknown" and left to human interpretation. Most often, these were cases where the dates were malformed, such as "4/301999", which is missing a/between the day and the year.

Once the system converted all of the dates that it could to a standard format, the system performed a check on the span of identified dates, to make sure that the difference between the earliest and latest date is less than 90 years. This check is motivated by the HIPAA requirement that ages over 89 be obfuscated. If the ages in a document are already annotated, identifying the ones that are over 89 is trivial; our system changed all the ages over 90 to "90" and left the other ages alone. However, if a patient is over 89 and the file lists their birth date and the current date, it would be easy to infer their actual age if all the dates are shifted consistently. Therefore, if the span of dates in a document is over 90 years, the system identified the earliest dates and shifted them more years into the future than the other dates in the records.

Finally, the system randomly selected the number of years and days that it would use to shift the dates in the record. We limited the number of years that dates could be shifted forward to 65 years plus or minus a maximum of 20 years for the 2014 i2b2/UTHealth corpus. The number of days we shifted the dates could be positive or negative, and we calculated this number by determining the maximum number of days the dates could be shifted in a direction without changing the season of any of the dates, then selecting a random number between one and the maximum. The day shift also acted as the month shift: if we shifted all the days back by seven, October 1st would become September 24th.

However, this method of date shifting also introduces its own potential sources for error. Consider a medical record that contains the statement "he will schedule a follow-up for January". The medical implication of this phrase is changed drastically if the record is dated December 24th (implying that the patient needs to be seen mere weeks, or even days, after that visit) or January 3rd (implying that the patient is anticipated to be fine for another year). If our date-shifting algorithm keeps the date on the record to be sometime in December, then "January" can be unchanged without drastically changing the implied medical condition of the patient. However, if "December 24th" becomes "January 9th", and the mention of the "follow-up in January" remains unchanged, then the patient's implied condition goes from being in need of close monitoring to safe for another year.

Any isolated mention of a month has the same problem. Our system addressed this problem by assigning a hidden day-of-the-month variable to the unanchored

months and giving that variable the value of "15". Then it applied the normal date-shifting algorithm, and the month would shift to the next one if the day-shift pushed it into the next month. So, given the above example with "December 24" and the following "January", if the date-shift value was set to 17 then the new dates would be "January 8" and "February", because the "hidden" 15th of January would become February 1st. If the date-shift was less than 17, then January would remain January. However, any forward date-shift less than seven would mean that the December date would remain in December, so keeping "January" unchanged would be the appropriate action.

Our method for dealing with unanchored months is not foolproof and can add errors into the surrogate text. For example, if the original dates were "August 2" and "the following September", if the dates were shifted forward by 16 days, the results would be "August 18" and "the following October", due to the hidden variable "15". An ideal approach would be to implement a system that uses the surrounding context to infer the year in which the unanchored month took place. Unfortunately, the task of assigning calendar dates to temporal phrases in medical records is no mean feat. The 2012 i2b2 NLP shared task focused on temporality in clinical texts, and the best-performing system had a 0.9 f-measure for identifying temporal expressions, but only achieved 0.73 for accuracy of the attribute values, which included the day, month, and year for identified dates [26].

27.5.4 HIPAA Category 18: Other Potential Identifiers

Additionally, we had to address information in the documents that did not fall into the PHI categories we described above, but that still had the potential to compromise information about a patient. Some EMRs contain references to specific events that the patient took part in, such as "injured during Superstorm Sandy" or "enrolled in HEART-FAB study". This type of information we usually removed entirely, or modified to be less identifiable. "Injured during Superstorm Sandy" could have become "Injured during last week's thunderstorm".

27.5.4.1 Professions

As previously discussed, we included patient's professions in the expanded set of PHI categories. The algorithm we used to generate surrogates for professions was relatively simple: we either selected a new, random profession from a pre-compiled list, or we used a surrogate that had already been selected for the same PHI. From a software perspective, this was easy to implement, but viewing professions as PHI posed a challenge that ultimately had to be handled through human intervention.

A person's profession can influence their health risks and outcomes. For example, people whose occupations involve construction or other building-related work, such as electricians and plumbers, are more likely to be exposed to asbestos,

and therefore those workers are more at risk for mesothelioma than others.[5] The relationship between medical outcomes and professions led us to attempt to assign new professions with roughly the same types of medical risks, thus preserving anonymity without entirely losing potential medical causalities. Similarly, the job held by a patient's relative occasionally impacted the medical record. For example, a record might state "Patient's daughter is a registered nurse who will help him monitor his blood pressure and blood sugar." Changing "registered nurse" to "lawyer" or another non-medical profession would make the sentence less logical, while supplying "doctor" or "medical aide" as a replacement would maintain the logic of the narrative. Our surrogate system did not implement a job hierarchy, and so we did this portion of the surrogate generation by hand where necessary.

27.6 Errors Introduced by Surrogate PHI

We made every effort to maintain the continuity of patient's medical histories so that researchers could still use the files for medical NLP. However our prioritization of patient privacy required us to make decisions that fundamentally changed the nature of the corpus.

By shifting each patient's records by different intervals and by randomizing the locations in the text, we rendered the documents useless for epidemiological studies, as researchers cannot use the data to infer trends based on shared locations and points in time. Errors in date-shifting can lead to creating mistakes in patients' own timelines, and if the system mistakenly generates different surrogates for entities that are co-referent, information about the patient is again lost. Similarly, substituting professions, even ones of similar types, may remove relevant job-related risk factors. Finally, any modifications to a text can result in unrealistic narratives in the form of incorrect determiners, verb tenses, and other grammatical errors.

Having a human review and edit the output can help reduce some of these errors, such as lost co-references and grammatical issues, but this is a time-consuming process and may not be possible for most research teams focused on de-identification.

27.7 Relationship Between De-identification and Surrogate Generation

The task of a de-identification algorithm should be to identify PHI, not to obliterate it. This is because obliteration (e.g., by replacement by a marker such as "[**

[5] http://www.asbestos.com/occupations/.

NAME 33 **]") destroys information about the original format of that PHI. This information is crucial for generating realistic surrogates.

Furthermore, a surrogate generation system may be able to do a better job in generating surrogates if it has access to the features, patterns, and dictionary memberships that the de-identification system used to detect the PHI. Otherwise, it may need to re-do some of the work already accomplished by the de-identifier. Our experience also shows that determining accurate co-references in the original data would be a good step toward creating consistent surrogates. We have described some fairly ad hoc methods for doing this, but a more general purpose method that itself uses machine learning techniques to build models that identify co-reference might be more effective. It does seem that, contrary to our initial expectations, it is not necessarily optimal to factor de-identification and surrogate generation into separate sequential tasks, connected only by a narrow stream of PHI annotations.

27.8 Conclusion

In this chapter we have presented some methods and algorithms for approaching the task of realistic surrogate generation for narrative EMRs, and discuss some of the specific approaches we used when creating the de-identified dataset for the 2014 i2b2/UTHealth NLP shared tasks.

In the United States, HIPAA regulations require that researchers can only release medical records for research purposes if each patient has given consent or if the records have been de-identified in order to protect patient privacy. By removing authentic PHI and replacing them with realistic surrogates, we attempt to maintain the narrative structure of the original records. Our method for surrogate generation tackles PHI categories one at a time, and follows different mechanisms for generating appropriate surrogates for each kind. It respects co-reference, maintains ambiguities that naturally exist in the original data, but does not resolve them. It aims to minimize the ambiguities that appear as artifacts of the process. Our experiences show that while automated methods provide a good start at surrogate generation, manual review and human intervention makes surrogates natural. Nonetheless, a general method for surrogate generation comes with a cost—it can invalidate the data for some purposes. In our case, the date shifting choices make epidemiology research on this data untenable.

Future work in this area will need to focus on both identifying text that falls under HIPAA category 18 and developing reasonable surrogates for that information without removing important information about each patient's health. A resource that groups occupations by similar health risks could potentially help with this problem. Improved systems for detecting co-reference, especially ones that can account for misspellings and nicknames, would also benefit this area of research.

Acknowledgements This project was funded by NIH NLM 2U54LM008748 PI: Isaac Kohane, and by NIH NLM 5R13LM011411 PI: Ozlem Uzuner.

References

1. Berman, J.J.: Concept-match medical data scrubbing. How pathology text can be used in research. Arch. Pathol. Lab. Med. **127**(6), 680–6 (2003)
2. Chakaravarthy, V.T., Gupta, H., Roy, P., Mohania, M.K.: Efficient techniques for document sanitization. In: Proceedings of the 17th ACM conference on Information and knowledge management, pp. 843–852 (2008)
3. Clifford, G.D., Scott, D.J., Villarroel, M.: User Guide and Documentation for the MIMIC II Database, database version 2.6. Available online: https://mimic.physionet.org/UserGuide/UserGuide.html (2012)
4. Deleger, L., Lingren, T., Ni, Y., Kaiser, M., Stoutenborough, L., Marsolo, K., Kouril, M., Molnar, K., Solti, I.: Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. J. Biomed. Inform. Aug;50:173–83 (2014). doi: 10.1016/j.jbi.2014.01.014
5. Douglass M.M.: Computer-assisted de-identification of free-text nursing notes. MEng thesis, Massachusetts Institute of Technology (2005)
6. Douglass M.M, Clifford, G.D., Reisner, A., Moody, G.B., Mark, R.G.: Computer-assisted deidentification of free text in the MIMIC II database. Comput. Cardiol. **31**, 341–344 (2004)
7. El Emam, K., Buckeridge, D., Tamblyn, R., Neisa, A., Jonker, E., Verma, A.: The re-identification risk of Canadians from longitudinal demographics. BMC Med. Inform. Decis. Mak. **11**, 46 (2011)
8. Gardner, J., Xiong, L.: An integrated framework for de-identifying unstructured medical data. Data Knowl. Eng. **68**(12), 1441–1451 (2009)
9. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: PhysioBank, PhysioToolkit, and Physionet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215-e220 (June 13, 2000). http://circ.ahajournals.org/cgi/content/full/101/23/e215
10. Golle, P.: Revisiting the uniqueness of simple demographics in the US population. In: Workshop on Privacy in the Electronic Society (2006)
11. Gupta, D., Saul, M., Gilbertson, J.: Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am. J. Clin. Pathol. **121**(2), 176–186 (2004)
12. HHS (Department of Health and Human Services). Standards for Privacy of Individually Identifiable Health Information, 45 CFR Parts 160 and 164. December 3, 2002 Revised April 3, 2003. Available from: http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/introdution.html
13. Jiang, W., Murugesan, M., Clifton, C., Si, L.: t-Plausibility: semantic preserving text sanitization. In: 2009 International Conference on Computational Science and Engineering (CSE), pp. 68–75 (2009). doi:10.1109/CSE.2009.353
14. Kumar, V., Stubbs, A., Shaw, S., Uzuner, O.: Creation of a new longitudinal corpus of clinical narratives. J. Biomed. Inform. 2015.
15. Kushida, C.A., Nichols, D.A., Jadrnicek, R., Miller, R., Walsh, J.K., Griffin, K.: Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med. Care **50**, S82–S101 (2012)
16. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann, San Francisco (2001)
17. Lafky, D.: The Safe Harbor method of de-identification: an empirical test. Fourth National HIPAA Summit West. http://www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf (2010)
18. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Doklady Akademii Nauk SSSR. **163**(4), 845–848 (1965) [Russian]. English translation in Sov. Phys. Dokl. **10**(8), 707–710 (1966)

19. Li, M., Carrell, D., Aberdeen, J., Hirschman, L., Malin, B.: De-identification of clinical narratives through writing complexity measures. Int. J. Med. Inform. **83**(10), 750–767 (2014)
20. McMurry, A.J., Fitch, B., Savova, G., Kohane, I.S., Reis, B.Y.: Improved de-identification of physician notes through integrative modeling of both public and private medical text. BMC Med. Inform. Decis. Mak. **13**, 112 (2013). doi:10.1186/1472-6947-13-112
21. Meystre, S., Friedlin, F., South, B., Shen, S., Samore, M.: Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med. Res. Methodol. **10**, 70 (2010)
22. Meystre, S., Shen, S., Hofmann, D., Gundlapalli, A.: Can physicians recognize their own patients in de-identified notes? Stud. Health Technol. Inform. Stud Health Technol Inform. 2014;205:778–82
23. Neamatullah, I., Douglass, M., Lehman, L.-W., Reisner, A., Villarroel, M., Long, W., Szolovits, P., Moody, G., Mark, R., Clifford, G.: Automated de-identification of free-text medical records. BMC Med. Inform. Decis. Mak. **8**, 32 (2008)
24. Stubbs, A., Kotfila, C., Uzuner, Ö.: Automated systems for the de-identification of longitudinal clinical narratives. J Biomed Inform. 2015 Jul 28. pii: S1532-0464(15)00117-3. doi: 10.1016/j.jbi.2015.06.007
25. Stubbs, A., Uzuner, Ö.: Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus J Biomed Inform. 2015 Aug 28. pii: S1532-0464(15)00182-3. doi: 10.1016/j.jbi.2015.07.020
26. Sun, W., Rumshishky, A., Uzuner, Ö.: Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J. Am. Med. Inform. Assoc. Published Online First 5 April 2013
27. Sweeney, L.: Replacing personally-identifying information in medical records, the scrub system. In: Cimino, J.J. (ed.) Proceedings, Journal of the American Medical Informatics Association, pp. 333–337. Hanley and Belfus, Washington (1996)
28. Sweeney, L.: Uniqueness of Simple Demographics in the U.S. Population. Carnegie Mellon University, School of Computer Science, Data Privacy Laboratory, Technical Report LIDAP-WP4. Pittsburgh (2000)
29. Uzuner, Ö., Luo, Y., Szolovits, P.: Evaluating the state-of-the-art in automatic de-identification. J. Am. Med. Inform. Assoc. **14**(5), 550–563 (2007)
30. Uzuner, Ö., Stubbs, A., Xu, H., co-chairs.: "Data Release and Call for Participation: 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data". https://www.i2b2.org/NLP/HeartDisease/
31. Yeniterzi, R., Aberdeen, J., Bayer, S., Wellner, B., Hirschman, L., Malin, B.: Effects of personal identifier resynthesis on clinical text de-identification. J. Am. Med. Inform. Assoc. **17**, 159–168 (2010)