Recommendation-based De-Identification

A Practical Systems Approach towards De-identification of Unstructured Text in Healthcare

Varun Bhagwan
Healthcare Informatics
IBM Almaden Research Center
San Jose, CA, USA

Tyrone Grandison
Healthcare Transformation
IBM Research
Yorktown Heights, NY, USA

Carlos Maltzahn
Computer Science
Univ. of California, Santa Cruz
Santa Cruz, CA, USA

Abstract — **In healthcare, de-identification is fast becoming a service that is indispensable when medical data needs to be used for research and secondary use purposes. Currently, this process is done either manually, by human agent, or by an automated software algorithm. Both approaches have shortcomings. Here, we introduce a framework for enhancing the outcome of the current modes of executing a de-identification service. This paper presents the steps taken in conceiving and building a privacy framework and tool that improves the service of de-identification. Further, we test the usefulness and applicability of this system through a study with HIPAA-trained experts.**

Keywords: privacy, risk and compliance, business process

I. Introduction

De-identification, hiding information that can lead to identification of an individual, is critical to unlocking the potential of primary care health data [20]. Currently, de-identification is done either through software algorithms or manually [5, 10, 15-19]. Unfortunately, both methods have significant limitations. Computer algorithms that discover and de-identify personally identifiable information (PII) have well-known shortcomings with respect to re-identification risk and usability reduction [24, 27]. These algorithms are generally not re-usable in other contexts from the one they were designed for [10]. Human-centric efforts that manually identify and transform sensitive content have proven to be inefficient and infeasible, especially for large datasets, and produce resultant sets with a high proportion of errors [5, 10, 12]. While humans tend to be more precise with the items they identify, automated de-identification algorithms are more scalable and can identify a larger number of possible candidates for transformation.

Human efforts to manually identify PII fall under a sub-class of what is known in computer science literature as Human Computation – a mechanism to leverage human abilities for solving complex computational tasks. From an economic standpoint, human computation has led to the creation of web-based markets for "crowdsourcing" diverse tasks, such as image labeling, transcription, editing, composition, etc. As reflected on Amazon's Mechanical Turks (MTurk) [21], these tasks are usually short, require little to no expertise on the part of the human worker, and "cost" a few cents per task. By contrast, the task of identifying PII in unstructured text usually requires significant effort, skill and is relatively expensive.

The effort required leads to the recognition that de-identification of unstructured text is difficult, at best, and it is only possible to attain a de-identified data set that is "private to the best of current known knowledge". The de facto standard is the use of software agents. In this paper, we explore a hybrid approach, where software and human agents are formally integrated and their individual advantages leveraged.

As both approaches to de-identification use the rules stated in the legal mandates in the healthcare sector, we first review the legislative framework in place that drive both software and human agents when they must execute a de-identification service. Then we present the background details of this emerging field and describe the relevant work in the space. After those discussions, we present the system's design considerations, introduce the Recommendation-based De-Identification (Re-DId) system and framework, and report the experimental results and lessons learned.

II. Legislative Mandate

In the USA, the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [29], enacted in 1996, and the associated Health Information Technology for Economic and Clinical Health (HITECH) Act passed in 2009 [30], define safeguards for Protected Health Information (PHI) that is held or transmitted by a *covered entity*[1] and their *business associates*[2].

The Privacy Rule defines two ways for a covered entity to determine that health information is de-identified: the Statistical approach and the Safe Harbor approach. The statistical approach requires that a qualified statistical or scientific expert conclude that the risk the information could be used alone, or in combination with other reasonably available information, to identify the subject is very small. Due to the qualitative nature of "*risk is very small*", this approach is difficult to pursue, especially when dealing with unstructured (textual) medical records. The Safe Harbor approach allows a covered entity to consider data to be de-

[1] Covered Entities are: 1) health care providers (which includes doctors, clinics, nursing homes, pharmacies, dentists, chiropractors etc.), 2) health plans (which includes health insurance companies, HMOs, company health plans, Medicare, Medicaid etc.), and 3) health care clearinghouses (which includes entities that process non- standard health information they receive from another entity into a standard, or vice versa)

[2] Business Associates are entities (persons or organizations) other than a member of a covered entity's workforce, who perform functions or activities on behalf of covered entity.

identified if it removes 18 types of identifiers and there is no actual knowledge that the remaining data could be used to identify an individual, either alone or in combination with other information. The HIPAA 18 Identifiers, according to 45 CFR §164.514(b)(2)(i), are the following data types for the individual, their relatives, their employers, or household members of the individual:

1. Names.
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP Code, and their equivalent geographical codes, 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.
   b. The initial three digits of a ZIP Code for all such geographic units containing 20,000 or fewer people are 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. Facsimile 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 record locators (URLs).
15. Internet protocol (IP) address numbers.
16. Biometric identifiers, including fingerprints and voiceprints.
17. Full-face photographic images and any comparable images.
18. Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification.

The Safe Harbor approach is the method-of-choice for executing compliance with the HIPAA Privacy Rule, primarily because it is the most direct and well-defined way of doing so. Thus, the Safe-Harbor approach currently constitutes the guiding principles for de-identification in healthcare (in the United States).

III. Background

This work is grounded in the fields of Automated software De-identification and Human Computation.

Automated De-identification

Numerous systems and approaches have been developed or proposed over the years for de-identifying medical text [5, 10, 15–19, 25, 27]. Automated PHI identification algorithms generally fall into two categories: pattern matching (rule-based) and machine learning. Some systems combine both approaches to identify different types of PHI, but the vast majority of de-identification systems rely on pattern matching, rules and dictionaries.

A rule-based approach uses a high-level description of the format of a token, e.g. a telephone is a sequence of three numbers followed by a hyphen followed by three numbers followed by a hyphen followed by four numbers, to specify the patterns to be identified in a piece of text. The advantage of rule-based approaches are that they require little to no training data, and can be easily and quickly edited for improved performance through addition and re-ordering of rules, dictionaries, regular expressions, etc. [10]. Unfortunately, they require a large amount of effort from skilled workers to design the complex rules for spotting the different types of PHI. Additionally, they are heavily customized to the dataset they are built against, and usually require a significant amount of re-work when faced with different datasets. Pattern matching also suffers from the need to enumerate all possible formats for each PHI type, e.g. naming conventions, date conventions, etc., which can get ponderous over time.

Typical Machine Learning (ML) approaches “learn” characteristics of data by way of pre-labeled examples. Current ML approaches for de-identification - typically Support Vector Machines (SVM) [2], Conditional Random Fields (CRF) [9], Decision Trees [14], etc. - are of the supervised nature. Their great disadvantage is that a large corpus of annotated content is required to train (and subsequently test) them. Annotating a large corpus of data is not easy [5, 31] - it takes an inordinate amount of time and resources, and requires significant involvement of domain experts. Invariably, almost all of the machine learning approaches tend to add some form of pattern matching to extract features or to detect certain PHI types, such as social security numbers, phone numbers, etc. While machine learning algorithms hold the promise of being able to automatically learn complex PHI patterns without significant domain expertise and their speed doesn’t decrease over time, trying to determine why the algorithm gave a specific result is challenging. Moreover, when applied to a new dataset, machine learning algorithms require additional annotated training data.

Rule-based methods perform better with PHI that has limited support in text (i.e. when there is not sufficient training data for machine learning approaches to be effective), but are more difficult to generalize. Machine Learning approaches perform better when dealing with PHI that may not be mentioned in dictionaries or codified by rules. Currently, the majority of automatic de-identification systems target some types of identifying information, not the complete set of 18 classes of PHI as laid out by HIPAA (section II) [10].

There is also the k-anonymity family of de-identification techniques [25]. In addition to the fact that these are techniques suited for structured records, they suffer from a basic flaw, i.e. the fundamental assumption that they make is questionable. Their foundation is that records with identifiers can be split into disjoint sets of quasi-identifiers (attributes which can supposedly be generalized and thus released), sensitive attributes (those that are not released) and non-private data. In reality, there are a number of attacks that enable sensitive information to be extracted even after these techniques have been applied [27].

The current set of de-identification algorithms are normally evaluated only over a specific type of a dataset, e.g. nursing notes, pathology reports, consult notes, echo reports, etc., instead of over heterogeneous data. It should be noted that in just the U.S. Department of Veterans Affairs (VA), 80% of all clinical documents are spread across 100 different clinical document types [10]. Moreover, in a majority of the cases, the PHI identification algorithm was aided by additional external content, such as patient demographic data, account numbers, etc. Given the large number of different clinical document types, and their variability across medical institutions, it can be quickly surmised that none of the automated de-identification approaches will reliably and consistently spot all different PHI types in accordance with HIPAA.

Human Computation

Human Computation [34] is a new and evolving research area in computer science that has gained prominence only in recent years [32], even though its usage in the computing context can be traced back to 1950s [28]. Its goal is to leverage human abilities for solving computational tasks for which no algorithmic solution exists. It does so by exploiting the difference in the abilities and costs between humans and algorithms for symbiotic human-computer interaction, one where the traditional roles of humans and computers are frequently reversed - the computer poses questions for the human to solve, and then collects and aggregates the solution(s). A variety of approaches can be utilized, usually dependent on a combination of the available skills and desired goals.

Characteristics of Human Computation requirements have been classified into the following six groups: Motivation, Quality Control, Aggregation, Human Skill, Process Order, and Task-Request Cardinality [23]. Process order refers to the order in which the task is carried out, and the building blocks are Computer (usually to farm out tasks), Worker (the human(s) carrying out the task), and the Requestor (the “owner” of the task). Representing a task through Process Order depicts the sequence in which the task will be executed, and by which block. The Task-Request Cardinality refers to the cardinality of a task to a worker. For instance, it can be One-to-One (one worker to one task), Many-to-One (many workers to one task), etc. Although many of these tasks require aggregation and agreement between a large number of humans (hundreds of thousands to even millions), our approach, presented in a later section, is reliant on a relatively fewer number of humans, albeit with a specific skillset - that of being able to identify PHI within medical text.

Human Computation is used in a variety of ways. On one hand, it is used to replace algorithmic approaches for problems where humans do a vastly superior job compared to state-of-the-art algorithms [33]. On the other hand, it is used to generate training data for algorithms and thus overcome the knowledge-acquisition/cold-start problem [1, 26]. Human Computation efforts where workers manually identify PHI in datasets and redact or transform sensitive content have proven to be inefficient and infeasible, especially for large datasets, and produce resultant sets with a high proportion of errors [5, 10, 12]. These errors are different than the ones found in automated de-identification approaches. Specifically, while humans tend to be more precise with the items they identify, they miss out on identifying a significant amount of sensitive data elements. This is the key insight exploited by Re-DId. We use automated de-identification to get coverage, and human computation to enhance precision.

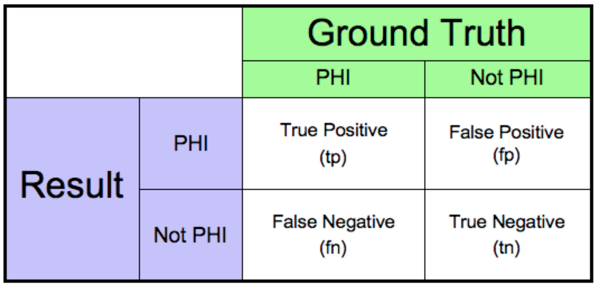
IV. Design Considerations

A system engineered to provide effective and measurable removal of PHI from medical data needs to demonstrate the following characteristics:

1. Handle heterogeneous data – this is necessary due to the wide variety of medical data across different hospitals, locations, specialties etc.
2. Provide minimal PHI exposure – to reduce risk stemming from litigation, fines and loss of image/reputation
3. Maintain usability of data for all existing and unforeseen uses – to further data-driven research in healthcare

As discussed in the prior section, no existing de-identification system meets the above requirements. The key challenge lies in the inability of computer algorithms to fully understand human-generated content. Humans are better at understanding unstructured content, but are unable to scale to be able to handle large datasets and overlook critical elements at times. These considerations were the driving factors behind the Re-DId system.

The primary goal of the system is to provide a mechanism to remove PII from datasets while maintaining usability. In the context of healthcare, the goal then is to build a system for effectively removing PHI from medical notes, while maintaining usability of said notes for medical



		Ground Truth	
		PHI	Not PHI
Result	PHI	True Positive (tp)	False Positive (fp)
	Not PHI	False Negative (fn)	True Negative (tn)

Figure 1: Ground Truth vs. Obtained Results

research and secondary uses. The output is a classification, wherein data elements are categorized as either PHI or not.

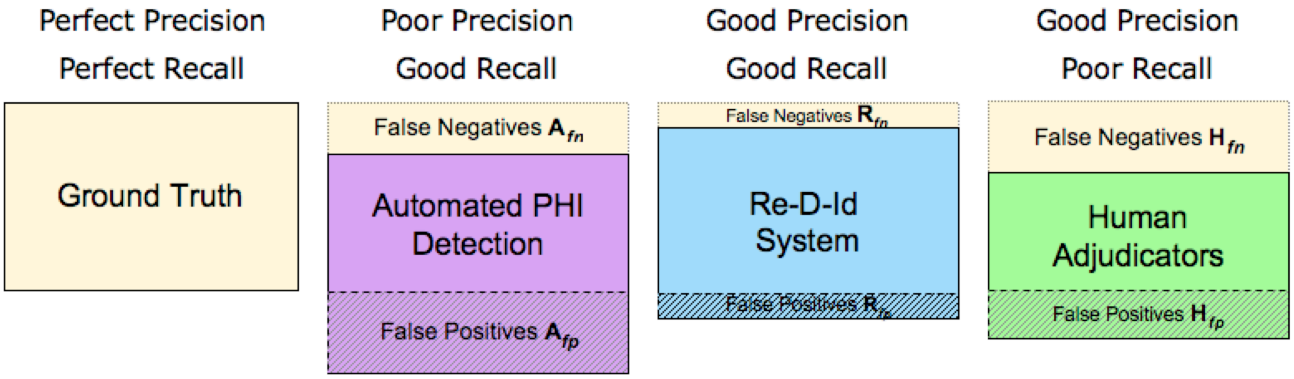


Figure 2: Relative Comparison using Precision and Recall

PHI elements not identified in a medical note are considered *false negatives (fn)*, while any non-PHI elements marked as PHI are considered *false positives (fp)* and all elements correctly identified as PHI are *true positives (tp)*. Leveraging terms from the text mining research community, we can define the concepts of *Precision* and *Recall* in terms of *fn*, *fp* and *tp*, where ”Ground Truth” is the actual classification of PHI, and ”Result” is the classification produced through a system (Figure 1).

Precision and Recall are defined as:

$$\text{Pr}\,ecision = \frac{tp}{tp + fp} \qquad \text{Re}\,call = \frac{tp}{tp + fn}$$

A system with a large number of false negatives is said to have *low Recall*, while one with a large number of false positives is said to have *low Precision*.

When manually identifying PHI in medical notes, humans are not only slow (compared to automated algorithms), but also very error prone [5, 4]. Specifically, they miss out on identifying a large amount of PHI, resulting in a high number of false negatives (where H_{fn} is the *fn* by Humans) when compared to the automated de-identification (where A_{fn} is the *fn* by Algorithms) approaches. The latter, due to their aggressive classification of data elements, misses out on far fewer PHI elements, and is significantly better at scaling.

Unfortunately, challenges still remain with the algorithmic approaches towards PHI de-identification [24, 10]. Indeed, when it comes to false positives, humans (where H_{fp} is the *fp* by Humans) have a much lower occurrence than automated algorithms (where A_{fp} is the *fp* by Algorithms). Intuitively, this makes sense, because humans are more adept at understanding text and language, while computers aren’t. For example, it is relatively trivial for a human to determine that the element ”*7/10*” in the note ”*patient c/o chest pain - 7/10, reduced to 2/10 after medicating with...*” is referring to the pain level. An aggressive algorithm, on the other hand, may mark this as a PHI element of type date.

Thus, when de-identifying medical text, there are distinct advantages and disadvantages of both humans and automated algorithms (Figure 2). Humans do not have as many false positives as automated de-identification algorithms, i.e. humans have better precision ($A_{fp} >> H_{fp}$), but miss out on a significant amount of PHI resulting in poor recall. By contrast, automated de-identification algorithms have a lower number of false negatives compared to humans, i.e. automated algorithms have better recall ($H_{fn} >> A_{fn}$), but they tend to over-redact, resulting in poor precision. In the Re-DId system, we exploit the complementary nature of human adjudication vs. automated de-identification.

V. Re-DId: System and Framework

The Recommendation-based De-Identification (Re-DId) approach (Figure 3) is comprised of three phases: first, execute an (off-the-shelf) automated de-identification algorithm on the chosen dataset (Phase 1); then use the output as source data for a recommendation engine whose function it is to surface items that were incorrectly identified and items that should have been de-identified (Phase 2); and finally present all these candidates to human workers for further adjudication (Phase 3). We recognize the need to aid the cognitive effort of human adjudication through analytical techniques. The resulting Re-DId system leverages existing algorithmic approaches and performs analytics to generate recommendations for PHI and non-PHI elements that are subsequently presented to a human worker for adjudication.

A. Phase 1: Automated De-identification Algorithm

In the automated de-identification phase, the free text medical documents to be redacted are fed to an arbitrary, state-of-the-art de-identification algorithm, along with any supplementary data required and or available. It should be noted that Re-DId is agnostic to the automated algorithm in use. We treat the algorithm essentially as out-of-the-box software, as it mimics usage in real world. Independent of the specific implementation details of the algorithm, we expect certain specific attributes in the output:

- PHI element – the term/phrase that is marked as PHI

- Span – the start/end offset within the record to accurately identify the PHI element
- Transformation [optional] – the transformation function performed on the PHI element (e.g., offsetting a date)
- Type [optional] – the type of the PHI element identified
- Confidence/Weight [optional] – a numeric value (normalized to between 0 and 1) that denotes the confidence or weight associated with the PHI element (higher value implies higher confidence).

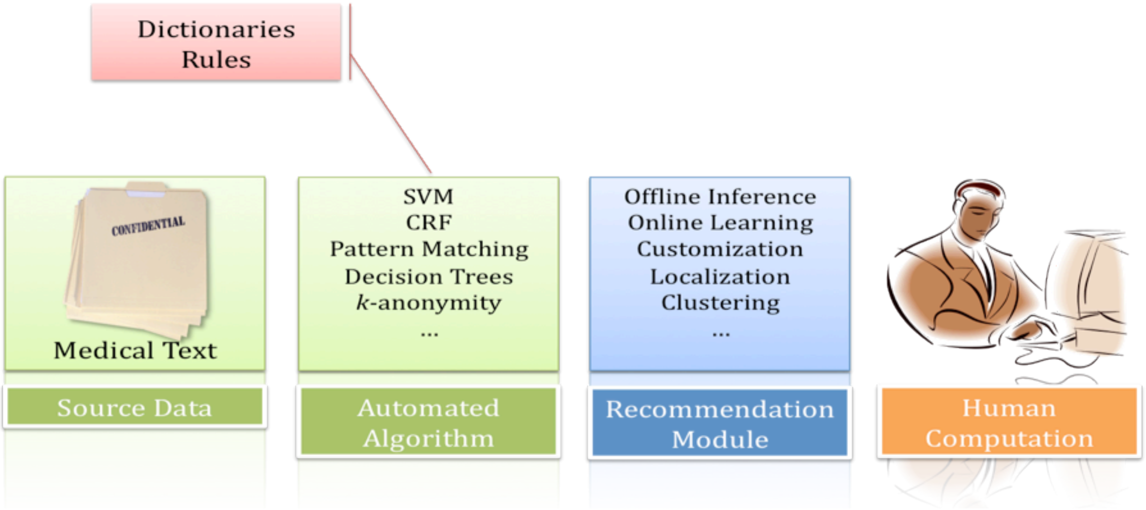


Figure 3: Re-DId System Overview

We should point out that transforming the output of existing de-identification algorithms to the above format is trivial, and therefore not a limitation by any means.

B. *Phase 2: Recommendation Module*

This phase takes the output of the previous phase (and the source files) and uses it to generate recommendations for PHI and non-PHI candidate sets. The techniques we have developed, presented below, are not exhaustive and we provide them only as a starting point for further research.

The *Customization* technique leverages domain-specific knowledge, e.g. dictionaries and ontologies for drugs, treatments, diseases etc., to generate recommendations. This helps to resolve any ambiguity in identifying data that is medically relevant, and thus potentially non-PHI. Additionally, automated de-identification algorithms are usually trained on specific types of data, e.g. patient-physician directories, hospital and location names, etc., and need to be re-trained in order to handle data that has different characteristics. We overcome this by enabling the requestor (i.e. the party that requests de-identification) to add location and or hospital specific data (provider directories, local taxonomies for disease and drug names, etc.). This technique improves both precision and recall.

The *Clustering* technique groups the data elements based on the entity-type they represent (e.g. dates, names, numbers, phone numbers, etc.), and surface the relevant types. This technique also leverages uncertainty, as in the data elements whose types are uncertain may be strong candidates for human adjudication. It is appropriate when automated de-identification algorithms of Phase 1 did not recognize PHI elements that did not match a syntactic structure (e.g. 8-digit medical record number), perhaps due to a typo (e.g. resulting in a 10-digit number). This technique primarily improves recall.

The *Offline Inference* technique directly leverages the PHI classification performed by the automated de-identification algorithm. This is done through analyzing the similarity between the remaining (un-redacted) data elements and the data elements labeled as PHI by the algorithm. The intuition being that there may be data elements that do not fit the syntactic structure the automated algorithm is looking for (e.g. due to a location-specific prefix being added), but are still PHI.

The *Online Learning* technique uses the adjudications that a human is making on a dataset, while they are doing it. The intuition here is that a data element consistently being adjudicated as PHI by the human worker is a strong indication that the same element in a similar context in a different note is a good candidate to be recommended as a PHI element if it occurs in the same dataset.

The final technique (*Fusion*) is a combination of all the prior techniques. Further to these techniques, the authors purport that there is a world of new and innovative analytics to be created by the broader research community that can be used to provide the human adjudicator with candidate annotations.

C. *Phase 3: Human Adjudication*

In this phase, human workers that have received adequate training on the HIPAA Privacy Rule [29] are employed to adjudicate on the PHI/non-PHI candidate sets generated in the prior phase. Although we do not describe the design, development and evaluation of the user-interface in this paper, the key characteristics desired of a representative interface are tabulated below:

- Simple – this is very important, as it is easy to clutter an interface with ”functionality”. It should be kept in mind that the consumers of this interface are not just computer scientists.

- Context – when adjudicating data elements, it is imperative to have context around the element prior to adjudication.
- Inline Lookup – for certain elements, it is useful to do a lookup, be it an Internet search, or a medical dictionary etc. Making such functionality available such that the participant doesn't need to exit the interface is important to a successful user experience.
- Interesting – this is highly subjective, but user studies are in general tedious activities. Keeping the participant engaged - for example, by displaying a summary ”score” of their adjudications upon reaching pre-determined milestones - is critical.

The Human Computation aspect of the Re-DId system works as follows. Adjudicators (human workers) are presented with the medical note as well as the recommendations for PHI/not-PHI candidates, and are asked to adjudicate each along the lines of:

- Yes - the data element is confirmed as belonging to the class (PHI/not-PHI) under consideration
- No - the data element is confirmed as not belonging to the class under consideration
- Unknown - the adjudicator is unsure, and the element is a good candidate for review by the Final Reviewer/domain expert

Finally, the adjudication results of multiple workers are aggregated and presented to the Final Reviewer in an optional step – or when the agreement amongst adjudicators is below a certain threshold. Optionally, the human adjudicators can also explicitly defer the most ambiguous of data elements to be adjudicated by a domain expert such as the Final Reviewer.

This small subset of the entire corpus is then reviewed and adjudicated by the Final Reviewer. The obvious advantage of this approach is that the human adjudicators in Phase 3 can be relatively unskilled (and thus lower paid) workers trained on the HIPAA 18 identifiers, while referring any ambiguous and or domain-specific items to the Final Reviewer. It is even conceivable to conduct Phase 3 on a system similar to Amazon's Mechanical Turk [21], should a means to provide and verify appropriate HIPAA training become feasible at some point in the future.

This 3-phase approach of the Re-DId system allows it to achieve the design objectives described in the prior section. While using a state-of-the-art automated de-identification algorithm ensures capturing a majority of PHI elements, the recommendation module coupled with human adjudication enables identification of any remaining PHI elements while correcting the *over-redaction* from phase 1. Thus, the Re-DId system is able to de-identify data from heterogeneous data sources while minimizing PHI exposure and maintaining usability.

VI. Experimental Study

The goal of our experimental study is multi-fold. First, to execute a de-identification algorithm on data it had been developed for, and demonstrate that there was room for improvement (by identification of false positives in the algorithmic output). Second, to realize this improvement via the Re-DId framework – by having human adjudicators rate the output of the de-identification algorithm. Third, to prove the validity and re-usability of the results of human adjudication by measuring the agreement amongst multiple adjudicators. Finally, to prove the assertion that off-the-shelf de-identification algorithms miss out on a significant amount of PHI elements when executed on data they have not been trained on.

A. *Experimental Setup*

For the experiments, the MIT Deid dataset is used as the input medical text that contains PHI. This dataset is a gold standard reference database of over 2600 nursing notes covering 148 patients (approximately 350,000 words) with about 1800 labeled instances of PHI. The dataset is available through a data-use agreement from http://www.physionet.org/physiotools/ deid/.

For the Automated PHI Detection Phase (Phase 1) of Re-DId, we executed the MIT Deid Algorithm [12] on the MIT Deid dataset. This algorithm was developed by the MIT team in conjunction with the MIT Deid dataset, and as such has as good performance as can be expected from a de-identification algorithm. However, as mentioned earlier, de-identification algorithms are heavily tuned for the datasets they are built against. The MIT Deid algorithm is no different. It uses a number of dictionaries that contain lists of PHI content on which the dataset itself is based.

The output of the MIT Deid Algorithm was a set of PHI elements for each record in the dataset. In the human computation step, the adjudicators considered each element and classified it as being PHI or not. Although each of the human participants had appropriate HIPAA training, as a preparatory step, they were briefed in the following manner:

- Review of the ”HIPAA 18” PHI identifiers
- An example of the interface and how to adjudicate
- Be conservative - when in doubt, err on the side of caution and mark the element as PHI
- They only need to consider the suggested PHI element(s), not actively scan for other elements (to keep the study focused and finish in a reasonable time with minimal user-fatigue)

The adjudicators saw over 300 records each, spread across 27 unique patients. The total number of PHI elements considered by the 6 participants was 4158.

When multiple raters (or classifiers) assign categorical ratings (classifications) to a number of items, a way to measure the inter-rater agreement is desirable. Fleiss' Kappa [8] is one such statistical measure, used to determine agreement between two or more raters. It calculates the degree of agreement over what would be expected by chance, and its value ranges from <= 0 for no agreement to 1 for complete agreement.

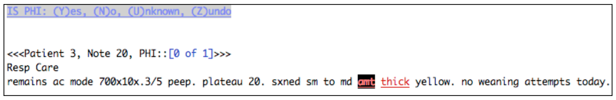
The Kappa, κ, is defined as:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

where $\overline{P}$ refers to the degree of agreement achieved, and $\overline{P}_e$ refers to the agreement that can be achieved by chance. Complete definitions of $\overline{P}$ and $\overline{P}_e$ can be found in [8]. So κ is $\overline{P} - \overline{P}_e$ which is the degree of agreement achieved above chance, normalized by $1 - \overline{P}_e$ which is the agreement attainable above chance.

B. *Human Computation and Interface Design*

Our goal for the Human Computation step was to provide a simple and clean interface to help the workers adjudicate on PHI elements one-by-one. To this end, we went through a few iterations of the user interface.



```
IS PHI: (Y)es, (N)o, (U)nknown, (Z)undo


<<<Patient 3, Note 20, PHI::[0 of 1]>>>
Resp Care
remains ac mode 700x10x.3/5 peep. plateau 20. sxned sm to md amt thick yellow. no weaning attempts today.
```

Figure 4: Study interface excerpt

In the final iteration, we incorporated the feedback received during the other phases of the development of the interface. The layout was as follows (Figure 4):

- We presented the entire medical note at a time, instead of a window around the PHI element in the note.
- All PHI presented in a given note was highlighted (red-underlined), but the PHI element being adjudicated was colored Red-on-Black. Adjudicators preferred seeing all upcoming PHI elements (rather than getting distracted when an upcoming obvious PHI element had no indication that it would be PHI).
- Each note was ”refreshed” upon adjudication, such that the Red-on-Black highlighting would automatically shift to the next PHI element. This made transitions appear seamless and increased the speed of adjudications.

C. *Experimental Results and Discussion*

In the context of our experimental setup, the first goal of this study was to evaluate if there was room for improvement for the MIT Deid Algorithm, i.e. find the number of false positives present in the resultset after running the MIT Deid Algorithm on the MIT Deid dataset.

The total number of false positive elements, as uncovered by study participants, was 856. In other words, 856 elements were not PHI, even though the algorithm marked them as such. The total number of elements confirmed as PHI by the algorithm was 3290, while 12 elements were marked as unknown. Thus, a total of 20.59% elements adjudicated as PHI by the MIT Deid algorithm were found to be false positives. The Fleiss’ Kappa Score for inter-rater agreement was 0.88, which implies near perfect agreement.

These results were extremely encouraging, and provide a strong validation of our approach. For one, we were able to establish that there is room for improvement in the results of an algorithm that was tuned and trained on that specific dataset. Second, we were able to establish the efficacy of computer-aided human computation, as the human workers used the algorithmic output for their adjudication and identified a large number of false positives. Third, with the strong agreement amongst the human workers, we were able to re-enforce the validity and usability of the results.

D. *Observations and Lessons Learned*

This study confirmed certain expected results, while surfacing unexpected questions, ideas and guidelines for subsequent studies. Select observations were:

1. The amount of time taken by each adjudicator varied greatly for a number of reasons: experience with medical notes, getting distracted when trying to understand the medical condition being discussed in the note, acronym familiarity, etc.
2. Somewhat surprisingly, even though a counter for the number of PHI elements in a medical note was provided (e.g., #3-of-8), the adjudicators ignored it and instead just focused on the highlighted PHI elements.
3. The “unknown” label was seldom used by any of the adjudicators. This is likely because they were told to be conservative, and mark an element as PHI when in doubt.
4. Although adjudicators were told to only consider one PHI element at a time (the one highlighted), they invariably looked ”around” that element to spot other PHI elements.
5. An ”undo” feature was a universally requested feature to correct inadvertent incorrect adjudication, even though the total number of such errors made by each adjudicator was at most in the low single digits.
6. A way to expand acronyms or conduct an inline web search for ambiguous and or medical terms should improve results, especially with relatively inexperienced adjudicators, and increase overall throughput.
7. Alternatively, or perhaps in addition, a separate option to ’refer’ the PHI element to a trained medical practitioner or specialist could prove to be useful – in lieu of labeling all uncertain elements as PHI, as was the case in the study.

E. *Further Evidence*

For a number of logistical and economic reasons, it is infeasible to train de-identification algorithms on all possible datasets that need to be de-identified. Yet, it is desired to de-identify all datasets prior to sharing them. The only choice then is to execute existing off-the-shelf de-identification algorithms on datasets that they have not been trained on. This, of course, presents the possibility of the leakage of PHI information in supposedly de-identified datasets.

In order to validate the claim that off-the-shelf de-identification algorithms do not work well on data they have not been trained on, we executed the MIT Deid Algorithm on the i2b2 dataset [31]. The dataset was transformed into the desired format for the algorithm and the results analyzed. By raw count, the output of the Deid algorithm contained only about 78% of the total elements identified as PHI in the i2b2 “Gold Standard” dataset. Further, this number depicted the *best-case* scenario as it assumed 100% precision, which was not the case (a quick visual scan of the results confirmed this).

Thus, we were able to establish that executing off-the-shelf de-identification algorithms on unfamiliar datasets will result in a significant number of *false negatives*, thereby exacerbating the need for a system that allows for refinement and correction of the algorithmic output.

VII. Conclusion

Identifying and removing PII from unstructured text is a challenging problem. We present a new class of systems that couple Human Computation with Algorithmic De-identification to keep sensitive (patient) data private, while sharing (medical) data pertinent for secondary use analysis. This approach is embodied in the Re-DId (Recommendation based De-identification) framework, which leverages the complementary nature of human computation and automated algorithmic de-identification to achieve good precision and ambiguity resolution from human computation, with good recall and scalability coming from the automated de-identification. We present the results of a study conducted with HIPAA-trained experts to demonstrate the viability and potential of the Re-DId approach – 21% false positives were identified, with near-perfect agreement across all adjudicators. We also demonstrate that executing off-the-shelf de-identification algorithms on datasets they have not been trained on is not sufficient, and requires additional rectification steps downstream. As future work, our goal is to further develop the recommendation module, create an evaluation framework for such coupled systems, and apply our privacy framework to other domains.

Acknowledgment

We wish to thank our colleagues Daniel Gruhl, Steve Welch, April Webster, Karen Brannon, Roxana Stanoi and Neal Lewis. We also wish to acknowledge the invaluable feedback and input from Dr. Joe Terdiman, Director, Division of Research, Kaiser Permanente.

References

[1] V. Bhagwan and C. Maltzahn, “JabberWocky: Crowd-sourcing metadata for files,” IEEE International Conference on Services Computing (2009), pp. 513–516.

[2] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” DATA MINING AND KNOWLEDGE DISCOVERY, 2 (1998), pp. 121–168.

[3] Canadian Government. Privacy Act. http://www.priv.gc.ca/legislation/02_ 07_01_01_e.cfm.

[4] D. Dorr, W. Phillips, S. Phansalkar, S. Sims, and J. Hurdle, “Assessing the difficulty and time cost of de-identification in clinical narratives,” In Methods of Information in Medicine (2006), pp. 246–252.

[5] M. Douglass, “Computer-assisted de-identification of free-text nursing notes,” Master’s thesis, Massachusetts Institute of Technology, 2005.

[6] Electronic Privacy Information Center. The Video Privacy Protection Act (VPPA). http://epic.org/privacy/vppa/, 2002.

[7] European Union. Data Protection Directive. http://ec.europa.eu/justice/policies/ privacy/index_en.htm.

[8] J. L. Fleiss, “Measuring nominal scale agreement among many raters,” Psychological Bulletin (1971), pp. 378–382.

[9] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” In Proc. 18th International Conf. on Machine Learning (2001), pp. 282–289.

[10] S. Meystre, F. F. FJ, B. South, S. Shen, M. and Samore, “Automatic de-identification of textual documents in the electronic health record: a review of recent research,” In BMC Medical Research Methodology (2010).

[11] National Institute of Standards and Technology. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII). http://csrc.nist.gov/publications/ nistpubs/800-122/sp800-122.pdf.

[12] I. Neamatullah, M. M. Douglass, L. wei H Lehman, A. Reisner, M. Villarroel, W. J. Long, P. Szolovits, G. B. Moody, R. G. Mark, and G. D. Clifford, “Automated de-identification of free-text medical records,” In BMC Medical Informatics and Decision Making (2008).

[13] Organisation for Economic Co-operation and Development. OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. http: //www.oecd.org/document/57/0,3746,en_ 2649_34255_1815186_1_1_1_1,00.html.

[14] J. R. Quinlan, “Induction of decision trees,” Machine Learning (1986), pp. 81–106.

[15] J. Gardner, L. Xiong, K. Li, J. J. Lu, “HIDE: Heterogeneous Information DE-identification (demo track),” In *12th International Conference on Extending Database Technology (EDBT)*, March, 2009

[16] K. El Emam, S. Jabbouri, S. Sams, Y. Drouet, M. Power, “Evaluating Common De-Identification Heuristics for Personal Health Information,” J Med Internet Res 2006;8(4):e28

[17] F. J. Friedlin, C. J. McDonald, “A Software Tool for Removing Patient Identifying Information from Clinical Documents,” J Am Med Inform Assoc. 2008 Sep-Oct; 15(5): pp. 601–610

[18] B. Wellner et al, “Rapidly Retargetable Approaches to De-identification in Medical Records,” J Am Med Inform Assoc 2007;14: pp. 564-573

[19] FP. Morrison, S. Sengupta, G. Hripcsak, “Using a pipeline to improve de-identification performance,” AMIA Annu Symp Proc. 2009 Nov 14;2009: pp. 447-451

[20] National Institute of Health, Research Repositories, Databases, and the HIPAA Privacy Rule. Available at: http://privacyruleandresearch.nih.gov/research_repositories.asp

[21] Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome.

[22] Hipaa training and compliance. http://www.hipaatraining.net/.

[23] A. J. Quinn, and B. B. Bederson, “Human computation: Charting the growth of a burgeoning field,” In CHI 2011 (Vancouver, Canada, May 7-12 2011).

[24] M. A. Rothstein, “Is de-identification sufficient to protect health privacy in research?,” The American Journal of Bioethics 10, 9 (Sept. 2010), 311.

[25] P. Samarati, and L. Sweeney, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,” In Technical Report, SRI Intl (March 1998).

[26] N. Seemakurty, J. Chu, L. von Ahn, and A. Tomasic, “Word sense disambiguation via human computation,” In Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP ’10, ACM (New York, NY, USA, 2010), pp. 60–63.

[27] S. Sundaresan, “Survey of privacy protection for medical data,” Master’s thesis, California State University at Los Angeles, 2008.

[28] A. Turing, “Computing machinery and intelligence,” Mind 59 (1950), pp. 433–460.

[29] U.S. Department of Health and Human Services. Standards for Privacy of Individually Identifiable Health Information; Final Rule. 45 CFR Parts 160 and 164.

[30] U.S. Dept. of Health and Human Services, Health Information Privacy. The Health Information Technology for Economic and Clinical Health (HITECH) Act. 2009.

[31] O. Uzuner, Y. Juo, and P. Szolovits, “Evaluating the state-of-the-art in automatic de-identification,” In Journal of American Medical Informatics Assoc. (2007), pp. 550–563.

[32] L. von Ahn, “Human computation,” PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2005. AAI3205378.

[33] L. von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, “reCAPTCHA: Human-based character recognition via web security measures,” Science. 2008 Sep 12;321(5895): pp.1465-8

[34] Wikipedia. Human-based computation. http://en.wikipedia.org/wiki/ Human-based_computation.

[35] Wikipedia. Personally identifiable information. http://en.wikipedia.org/wiki/ Personally_identifiable_information.

[36] Wikipedia. Anti-phishing act of 2005. http://en.wikipedia.org/wiki/ Anti-Phishing_Act_of_2005, 2005.

[37] J. R. Young, http://chronicle.com/article/To-Justify-Every-A-Some/128528/, 2011.