Published in final edited form as:

Methods Inf Med. 2016 August 05; 55(4): 356–364. doi:10.3414/ME15-01-0122.

Is the Juice Worth the Squeeze? Costs and Benefits of Multiple Human Annotators for Clinical Text De-identification

D. S. Carrell¹, D. J. Cronkite¹, B. A. Malin², J. S. Aberdeen³, and L. Hirschman³

¹Group Health Research Institute, Seattle, WA

²Vanderbilt University, Nashville, TN

³The MITRE Corporation, Bedford, MA

Summary

Background—Clinical text contains valuable information but must be de-identified before it can be used for secondary purposes. Accurate annotation of personally identifiable information (PII) is essential to the development of automated de-identification systems and to manual redaction of PII. Yet the accuracy of annotations may vary considerably across individual annotators and annotation is costly. As such, the marginal benefit of incorporating additional annotators has not been well characterized.

Objectives—This study models the costs and benefits of incorporating increasing numbers of independent human annotators to identify the instances of PII in a corpus. We used a corpus with gold standard annotations to evaluate the performance of teams of annotators of increasing size.

Methods—Four annotators independently identified PII in a 100-document corpus consisting of randomly selected clinical notes from Family Practice clinics in a large integrated health care system. These annotations were pooled and validated to generate a gold standard corpus for evaluation.

Results—Recall rates for all PII types ranged from 0.90 to 0.98 for individual annotators to 0.998 to 1.0 for teams of three, when measured against the gold standard. Median cost per PII instance discovered during corpus annotation ranged from \$0.71 for an individual annotator to \$377 for annotations discovered only by a fourth annotator.

Conclusions—Incorporating a second annotator into a PII annotation process reduces unredacted PII and improves the quality of annotations to 0.99 recall, yielding clear benefit at reasonable cost; the cost advantages of annotation teams larger than two diminish rapidly.

1. Introduction

Clinical text, such as physician progress notes, contains valuable information about the health conditions and treatment of patients whose care is documented in electronic health record (EHR) systems. This information can be reused in research if 1) the patient consents to such activities, 2) an Institutional Review Board (IRB) grants a waiver of such consent, or

3) the data is “de-identified” according to a regulatory standard. In the United States, the standard of greatest relevance is rooted in the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) of 1996. (1) Under the rule, patient-specific health information is designated as de-identified if there is no reasonable basis to believe such information can be used to identify an individual. There are two core approaches by which de-identification can be accomplished; the first is the certification by an “expert” who shows the risk of identification is small, while the second corresponds to a Safe Harbor whereby the organization removes values associated with 18 designated fields and confirms they have no knowledge that the data could be invoked for re-identification purposes. With respect to clinical text, the majority of healthcare organizations tend to apply the latter strategy because of its easy-to-interpret definition and straightforward implementation.

De-identifying clinical text according to the Safe Harbor model may be accomplished via a wide variety of manual and automated processes. (2) In the manual setting, humans detect, redact and/or replace personally identifiable information (PII) from each document of interest. In the automated or semi-automated setting, humans annotate documents in a representative sample and then use this sample to train machine learned classifiers to recognize patterns (e.g., grammars) indicating which text spans are PII, optionally supplementing classifiers with single- or multi-state rule-based logic, (3) This process can then be scaled up using software. There is evidence, based on challenge competitions held at the AMIA Annual Fall Symposium, suggesting that automated strategies to de-identification are the most scalable and generalizable. (4, 5) However, to apply such sophisticated methods (e.g., decision trees, conditional random fields, or support vector machines) (6) each institution must take the time to select and annotate its own local corpus. (7, 8) This is particularly the case because, while one might believe that a public synthetic corpus (e.g., i2b2 or Cincinnati) (5, 9) might be reusable, it has been shown that de-identification models based on such data underperform in practice, leaving un-redacted (and thus residual) PII instances even the best automated de-identification systems will overlook. (9, 10) There is thus strong motivation to develop approaches and techniques to assist humans in annotation of clinical text from their own institution. (11)(11) Additionally, there are efforts underway to develop novel de-identification strategies (e.g., replacing redacted PII instances with fictitious surrogates) to reduce the risk of exposure associated with residual PII. (12)

Fundamental to addressing these de-identification challenges is the availability of accurately annotated corpora. Outside the clinical domain, where patient privacy is not a consideration, numerous annotation studies have used crowdsourced annotation to explore questions of annotation accuracy, variability and cost. (13, 14) Crowdsourced annotation is not feasible in real clinical text because of the difficulty of obtaining patient consent and the need to preserve patient privacy under a waiver of consent. Knowledge of annotation performance and cost in the clinical domain is therefore relatively sparse. A known impediment to robust annotation in the clinical domain is the varying reliability of human annotators. Dorr and colleagues have shown that there is considerable variation across human annotators performing the same de-identification task. (15) It is therefore standard practice to invoke the assistance of multiple humans when creating gold standard annotations to be used for model training. (9, 16, 17) Yet manual annotation is expensive and time consuming. (15, 18–21) In one study annotators marked 4 annotations per minute in well-formatted

grammatically correct documents. (22) Annotation speed is likely slower in natural clinical text, which is often jargon-filled and written in telegraphic style.

2. Objectives

These previously reported findings motivated our investigation into the costs and benefits of increasingly aggressive attempts to develop high-quality gold standard annotations of PII. We focused our analysis on recall rates (i.e., the proportion of all PII in a corpus discovered through the annotation process) because avoiding PII disclosure is a paramount concern of IRBs responsible for assuring compliance with state, federal, and local privacy regulations and policy. This work addresses two closely related questions motivated by the need to manually annotate corpora for purposes of data sharing or training/updating automated de-identification systems: 1) How many independent human annotators are needed to identify all or nearly all PII in a clinical corpus? 2) What is the cost of improving PII annotation to achieve increasing levels of coverage of actual PII?

To answer these questions we conducted an experiment in which four annotators trained in medical records abstraction independently annotated PII instances in a sample of real outpatient clinical notes. We evaluated annotators' performance individually and in groups formed by combining individual annotators' results into hypothetical teams of two, three, or four annotators, thereby simulating the annotations that would be obtained by multi-person annotation processes. We also tracked time and personnel costs and used this information to estimate costs associated with increasingly aggressive annotation efforts.

3. Methods

3.1 Study corpus and annotation schema

The study corpus consisted of 100 actual clinical notes selected from patients' electronic charts, randomly selected from 4,438,028 Family Practice encounter notes authored by practitioners in Group Health clinics during a one-year period ending in October 2011. Prior research has shown that encounter notes are among the richest types of clinical text and among the most difficult to de-identify using scalable automated methods. (23) Short notes (<100 characters in length) were excluded because they were unlikely to contain PII. We extracted clinical notes from an Epic® Clarity® reporting database, and reassembled notes from multiple database records, as needed, to create verbatim copies of the complete Epic notes. As illustrated in Figure 1, clinical notes reassembled from the Clarity database version lack line breaks, paragraph breaks, and other white space that enhances a note's readability and, sometimes, its meaning (Figure 1, right pane).

Our PII annotation schema included 11 categories that encompass the 18 types of identifiers protected under HIPAA: patient names, patient ages, dates, social security numbers, telephone/fax numbers, medical record numbers, email addresses, Internet addresses (e.g., Web sites, computer network IP addresses), geographic locations (including addresses), room numbers (e.g., in hospitals), and a category for other identifiers not covered by the preceding categories (e.g., medical device serial numbers are "other identifiers"). In addition, we included categories for organization names (e.g., a hospital or place of

employment), practitioner names, and practitioner initials. We did this because many health care institutions prefer to remove such information before considering text to be de-identified. (24) Clinical text with correctly and incorrectly annotated PII spans is illustrated in Figure 1; the note shown has been fictionalized to preserve patient privacy but notes used in the study were actual notes from real patients.

3.2 Creation of the gold standard annotation set

Annotation tasks in this experiment were performed by four college-educated, experienced medical records abstractors who perform medical records abstraction as employees of a health research organization. Using fewer annotators would have reduced opportunities to observe variation in performance across abstractors; using more would have exceeded our available study budget. Each annotator received approximately two hours of standardized, in-person training explaining the guidelines for de-identification, the PII annotation schema, and the use of the annotation software to annotate documents. Each annotator was provided with a binder containing written annotation guidelines and formal definitions and illustrative examples of each PII type; they then used these guidelines to complete a practice exercise using real documents (not included in the experimental set). Annotators were instructed not to discuss their annotation work with any of the other annotators, but were encouraged to contact study staff with any questions concerning the annotation guidelines or their application to specific PII instances that might arise while reviewing documents. In responding to annotators' questions during document review, study staff restricted their answers to clarification of annotation guidelines, and avoided indicating whether a specific span of text should be annotated as PII. Annotators were instructed to take as much time as they felt they needed to identify all PII instances in each study document and to avoid fatigue by taking breaks and/or interleaving their annotation work for this study with unrelated tasks on other studies they were working on at the time. Each annotator was instructed to annotate documents in the corpus in the same order, which was determined by the annotation software interface. We retrospectively obtained each annotator's time spent and compensation rates from an administrative database.

We defined our gold standard as the combination of all PII instances that were annotated by any of the four annotators after review by a committee to exclude instances determined not to be PII. Annotations for which at least three annotators agreed exactly on span and PII type were incorporated into the gold standard without further review. All remaining instances marked by any annotator were reviewed by a committee of study team members to determine whether they satisfied annotation guidelines, and annotations that did not were excluded from the gold standard. This process assured that any text annotated by any annotator, including text that turned out not to be PII, was considered for inclusion. The committee planned to use a majority vote when decision by consensus was not possible, but this provision turned out not to be needed.

As with any gold standard, it is possible ours was imperfect. There is a very small chance that the study corpus contains additional PII instances not represented in our gold standard. Such imperfections would not, however, compromise our ability to assess the incremental costs and benefits of using additional human annotators when annotating PII in a corpus.

3.3 Evaluating annotation performance and cost

We evaluated the performance of each annotator (i.e., A1, A2, A3, and A4) by their individual recall rate, calculated by dividing the number of PII instances they annotated correctly by the total number of PII instances in the gold standard. A PII instance includes all of its component tokens (e.g., the three-token instance “Jane B Smith” correctly annotated in Figure 1), and each instance was determined to be either correctly or incorrectly annotated in its entirety—no partial credit was given. We considered a PII instance to be correctly annotated by an individual annotator if its span (i.e., the marks denoting the beginning and ending characters of a PII instance) matched that of the gold standard perfectly (see “Annotation correct” in Figure 1). We also counted as correct annotations that fully encompassed the gold standard span, even if too long (see “Annotation too long” in Figure 1) because such discrepancies do not reveal PII and, when annotation teams of size two or larger are used, they may be corrected by a standard reconciliation process. We counted as annotation errors PII instances that were overlooked entirely (see “Annotation missing” in Figure 1) and annotations that failed to include the entire gold standard span (see “Annotation too short” in Figure 1) because both types of errors could reveal PII. We ignored discrepancies in PII type (e.g., a blood pressure measurement marked as a date or vice versa), because such errors do not reveal PII; they only affect the information content of a de-identified corpus.

We also evaluated the annotation performance of simulated teams of two and three annotators. We did this by combining results of individual annotators to form all possible sets of two annotators and all possible sets of three annotators. This yielded six hypothetical teams of two annotators each (i.e., annotators A1+A2, A1+A3, A1+A4, A2+A3, A2+A4, and A3+A4) and four hypothetical teams of three annotators each (i.e., annotators A1+A2+A3, A1+A2+A4, A1+A3+A4, and A2+A3+A4). We defined a team’s annotations as the superset of correct annotations made by any of its individual members. For example, for the simulated two-person team of annotator A1 and annotator A2, the team was credited with annotating instances correctly marked by both annotators, instances marked correctly only by annotator A1, and instances marked correctly only by annotator A2. We calculated the recall rates for teams according to the same rules used in calculating individual recall rates (described above): number of PII instances correctly annotated by the team divided by the total number of PII instances in the gold standard.

We also summarized individual annotator and team annotation performance in terms of the minimum, median, and maximum number of PII instances overlooked by each annotator or team of annotators relative to the gold standard. We then used these metrics to calculate the expected cost and benefits of incorporating additional annotators into an annotation process.

To assess annotation costs we calculated the mean total personnel cost of employing annotators to annotate the 100-document corpus. We arrived at this cost by multiplying the mean number of hours annotators spent annotating by the mean personnel cost of the same annotators in current (2014) dollars. Personnel costs included employee wages, fringe benefit expenses, and indirect expenses such as office space and administrative overhead. We used an indirect expense rate of 0.60, applied to wages and fringe benefits, because 0.60 is typical for institutions receiving National Institutes of Health research funding. We used the

resulting per person cost figure to calculate the estimated cost per PII instance annotated, cost per document annotated, and cost per additional annotation discovered during gold standard creation by including additional annotators in the process. Time and cost estimates were based on data for three annotators because records for one of the annotators were not retained. This study was approved by the Group Health Human Subjects Review Committee.

4. Results

The 100 document corpus included notes authored by 88 unique practitioners for encounters with 100 unique patients totaling 37,010 words. Document word count ranged from 16 to 2,093 with a median of 373 (25th percentile = 154; 75th percentile = 506).

Ninety-four documents contained at least one instance of PII. Documents without PII tended to be short communications in which identifying information was unnecessary because it was not essential or could be inferred from the location of the document in the electronic chart. The entire corpus contained a total of 1,093 gold standard PII annotations of 11 types (Table 1). The most common forms of PII governed by HIPAA Safe Harbor were dates, patient names, and patient ages. The non-HIPAA PII types of organization names and practitioner names were the 2nd and 3rd most common forms of PII in this corpus, respectively. The set of gold standard PII annotations reflected contributions of all four annotators. As shown in Table 2, the vast majority of gold standard PII was discovered by all four annotators (915 instances or 83.7%). Declining proportions of gold standard annotations were contributed by only three annotators, only two annotators, and only one annotator (i.e., 14.2%, 1.6%, and 0.5%, respectively; Table 2). Notably, three of the four annotators made unique contributions to the gold standard, each contributing two PII instances not annotated by any other annotator.

Evaluation of recall rates relative to the gold standard for teams of one, two, and three annotators are presented in Table 3. Recall rates among individual annotators ranged from a low of 90.1% to a high of 97.8% with a median recall of 96.6%. These recall rates corresponded to counts of overlooked PII ranging from 24 to 108 instances (median 38). The recall rates among teams of two independent annotators ranged from 99.2% to 99.8% (median 99.5%).

The recall rates among teams of two independent annotators ranged from 99.2% to 99.8% (median 99.5%). These two-person rates corresponded to counts of overlooked PII in the corpus ranging from 2 to 9 (median 6). Recall rates for teams of three independent annotators ranged from 99.8% to 100% (median 99.8%). These three-person rates corresponded to counts of overlooked PII ranging from 0 to 2 (median 2).

Recall rates for each PII type by team size are shown in Table 4. Annotation performance for all PII types improved with team size with one exception: the five medical record numbers in the corpus were perfectly annotated by teams of all sizes. Teams of three annotators performed nearly perfectly across all PII types. The greatest variation in recall rates was observed among individual annotators. Variation in recall across all PII types for teams of two was substantially reduced compared to that observed for individuals.

Notably, one individual annotator failed to annotate any of the 90 patient ages in the corpus. Based on discussion with this annotator after data entry was complete, we determined that this oversight resulted from the annotator's misunderstanding of the annotation guidelines for patient ages. The guidelines distinguish between current ages and historical ages. A current age is illustrated in the statement "This 58 year old African American female presents today with abdominal pain." This age is PII because it can be used to infer the patient's current age and year of birth. In contrast, an historical age is illustrated in the statement "Her appendix was removed at age 22." Ages in historical references are not PII because they cannot be used to determine a patient's year of birth.

Mean time spent annotating the corpus by individual annotators was 16.38 hours. With a mean total personnel cost of \$46.03 per hour (in 2014 dollars) the mean per annotator cost of annotating the corpus was \$753.97 (16.38×46.03) and total annotation costs for creating the gold standard were \$3,015.88 (753.97×4). Given that the 100-document gold standard corpus contained 1,093 PII instances, this resulted in annotation costs of \$2.76 per PII instance and \$30.16 per document annotated.

Figure 2 shows the estimated cost of improving PII recall (i.e., reducing overlooked PII) in an annotation effort by adding additional independent annotators to the annotation task. Teams of two annotators discovered a median of 31 additional PII instances in the corpus (compared to the median of individual annotators) at an overall cost of \$24.32 per instance discovered ($\$753.97 / 31$). Increasing the annotation team size from two to three resulted in the discovery of an additional 4 instances of PII at a cost of almost \$188.49 per instance discovered ($\$753.97 / 4$). The "team" of four annotators whose annotations defined the gold standard discovered an additional two instances of PII compared to the median results for teams of three annotators. Though this is the only four-annotator team that annotated the corpus, we used their results to estimate the cost of adding a fourth annotator to be almost \$377 per additional instance annotated ($\$753.97/2$; see Figure 2).

5. Discussion

This study demonstrates that using multiple independent annotators to annotate a corpus for PII substantially reduces the instances of unredacted PII and therefore improves the quality of annotations. It also confirms that annotations produced by teams of multiple annotators exhibit substantially less variability than those produced by individual annotators, thereby eliminating much of the unpredictability associated with solo-annotator efforts. However, the marginal benefit of adding additional annotators after the second annotator declines sharply. There is a sharp increase in the overall costs associated with achieving incremental gains in completeness of annotations by adding more annotators.

Despite our efforts to provide all abstractors adequate training, one annotator (A4) misunderstood the guidelines, resulting in numerous failures to annotate patient ages. Though unfortunate, this outcome is a powerful reminder that any process that relies on a single mechanism, whether it is a person or an algorithm, harbors a risk of significant failure. Further, such failures may be difficult to detect; annotator A4's performance for PII types other than age equaled or surpassed that of other annotators. It is also instructive that

using teams of two annotators largely eliminated the annotation failures for patient age due to one annotator's errors.

Adding more annotators to an annotation team produces the dual benefit of improving overall recall and reducing variability. The observed variability in recall among the four individual annotators (range: 90.1% – 97.8%) is similar to that reported by Mayer and colleagues in a comparable annotation task (94%–96%). (18) Douglass and colleagues have reported somewhat lower recall rates. (25) Median recall among teams of two annotators was very high (99.5%) and continued to improve with the addition of a third annotator (99.8%). In addition, the predictability of the annotation result increases as the number of annotators on a team increases. Using the number of overlooked PII as a performance metric, the *worst-performing* team of three annotators performed at least as well as the *best performing* team of two annotators. Similarly, the *worst performing* team of two annotators performed at least as well as the *best performing* individual annotator.

We did not formally assess the potential loss of information caused by excessively long annotations. Such errors do not compromise patient privacy, but may result in information loss (e.g., if clinically relevant information is redacted as part of an adjacent PII annotation that mistakenly includes it). The risk of such information loss is expected to increase with the number of annotators independently annotating a corpus. In this experiment such annotation errors were rare, and in no case did they result in loss or distortion of clinical information. Nevertheless, trade-offs between protecting patient privacy and preserving information should be considered when multiple annotators are used to identify PII.

One implication of this study with respect to gold standard creation is that even the addition of a fourth annotator is likely to yield improvements, though the size of those improvements may be quite small. Three of the four annotators each contributed two PII instances to the gold standard set that their counterparts overlooked (Table 2 rows 1.1, 1.2, and 1.3). Even the individual annotator with the worst overall recall rate (Table 3 annotator A4), who overlooked 108 PII instances (relative to the gold standard), contributed two instances of PII to the gold standard other annotators overlooked (Table 2 row 1.3).

While incorporating additional annotators in an annotation task increases performance and reduces variability in annotation results, these benefits come with a price. Cost per additional instance of PII discovered increases exponentially as the number of annotators is increased from one to four, costing \$2.76, \$24.32, \$188.49, and \$376.99, respectively. We believe most secondary users of clinical text and institutional review boards would agree that adding a second annotator to an annotation task is “juice worth the squeeze.” However, the high and linearly increasing cost of incorporating additional annotators (over \$750 per annotator of a 37,000-word corpus) yields diminishing returns on investment. Under some circumstances, such as preparing a corpus for public release under a data use agreement, the cost of using four—or even more—annotators may be considered money well spent. Under other circumstances the additional cost burden of attempting to improve annotation completeness in a corpus (or reduce overlooked PII in a de-identification task) will outweigh the expected benefits, at which point “the juice is not worth the squeeze.” Newly reported research applying game theory to questions of clinical text de-identification suggests less than perfect

redaction of PII may be sufficient in some situations, (26) highlighting the relevance of quantifying the costs associated with incremental improvement in annotation completeness.

A limitation of this study is that it uses clinical corpora from a medical specialty in a single institution, reviewed in a given (random) order, and it is possible that inter-personal variation in PII detection rates may vary depending on the style, content, and/or order of documents in the corpus being annotated. Requiring all annotators to review documents in the same order allowed us to remove order-related effects from the experiment, but prevented us from investigating such influence. Future work should investigate costs and variations in annotation performance when using corpora of substantially different origin, style, content and review order. Another important question is whether completeness of PII annotation (recall) is most efficiently achieved through increasing the quality of annotators or the quantity of annotators. Our experiments used highly skilled, well-compensated annotators and encouraged meticulousness. It is possible, however, that an approach using a much larger number of less-skilled annotators and emphasizing speed (throughput) could achieve comparable results at the same or lower cost, given appropriate methods of combining the multiple annotation inputs.

6. Conclusions

Well-annotated clinical text can greatly improve the value of a corpus to researchers and the institutions responsible for its proper handling. Incremental increases in the accuracy and costs of annotation are important considerations when planning to create manually annotated gold standard corpora or when de-identifying clinical text by manual annotation. Incorporating a second independent annotator into a PII annotation tasks substantially reduces unredacted PII and improves the quality of annotations, yielding clear benefit at reasonable cost. The marginal costs and benefits of annotation teams of three or larger are less clear and may depend on the motivation driving the annotation work and the expected costs associated with omitted annotations.

References

1. U.S. Department of Health and Human Services. Standards for Privacy of Individually Identifiable Health Information; Final Rule. Federal Register. 2002;53181–53273. [PubMed: 12180470]
2. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol. 2010; 10:70. Epub 2010/08/04. [PubMed: 20678228]
3. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Combining knowledge- and data-driven methods for de-identification of clinical narratives. J Biomed Inform. 2015; 58(Suppl):S53–S59. Epub 2015/07/27. [PubMed: 26210359]
4. Stubbs A, Kotfila C, Uzuner O. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform. 2015 Epub 2015/08/01.
5. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc. 2007; 14(5):550–563. Epub 2007/06/30. [PubMed: 17600094]
6. Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, et al. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assoc. 2007; 14(5):564–573. Epub 2007/06/30. [PubMed: 17600096]

7. Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. *J Am Med Inform Assoc.* 2007; 14(5):574–580. Epub 2007/09/08. [PubMed: 17823086]
8. Uzuner O, Sibanda TC, Luo Y, Szolovits P. A de-identifier for medical discharge summaries. *Artif Intell Med.* 2008; 42(1):13–35. Epub 2007/12/07. [PubMed: 18053696]
9. Deleger L, Lingren T, Ni Y, Kaiser M, Stoutenborough L, Marsolo K, et al. Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. *J Biomed Inform.* 2014; 50:173–183. Epub 2014/02/22. [PubMed: 24556292]
10. 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.* 2010; 17(2):159–168. Epub 2010/03/02. [PubMed: 20190058]
11. Hanauer D, Aberdeen J, Bayer S, Wellner B, Clark C, Zheng K, et al. Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. *Int J Med Inform.* 2013; 82(9):821–831. Epub 2013/05/07. [PubMed: 23643147]
12. Carrell D, Malin B, Aberdeen J, Bayer S, Clark C, Wellner B, et al. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *J Am Med Inform Assoc.* 2012 Epub 2012/07/10.
13. Khare R, Burger JD, Aberdeen JS, Tresner-Kirsch DW, Corrales TJ, Hirschman L, et al. Scaling drug indication curation through crowdsourcing. *Database : the journal of biological databases and curation.* 2015; 2015 Epub 2015/03/24.
14. Khare R, Good BM, Leaman R, Su AI, Lu Z. Crowdsourcing in biomedicine: challenges and opportunities. *Briefings in bioinformatics.* 2015 Epub 2015/04/19.
15. Dorr DA, Phillips WF, Phansalkar S, Sims SA, Hurdle JF. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods Inf Med.* 2006; 45(3):246–252. Epub 2006/05/11. [PubMed: 16685332]
16. Grouin C, Névél A. De-identification of clinical notes in French: towards a protocol for reference corpus development. *Journal of Biomedical Informatics.* 2014; 50(0):151–161. [PubMed: 24380818]
17. Velupillai S, Dalianis H, Hassel M, Nilsson GH. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. *Int J Med Inform.* 2009; 78(12):e19–e26. Epub 2009/06/02. [PubMed: 19482543]
18. Mayer, J.; Shen, S.; South, BR.; Meystre, S.; Friedlin, FJ.; Ray, WR., et al. Inductive creation of an annotation schema and a reference standard for de-identification of VA electronic clinical notes; *AMIA Annu Symp Proc*; 2009. p. 416-420. Epub 2009/01/01
19. Jamison, EK.; Gurevych, I. Pacific Asia Conference on Language, Information and Computation. Phuket, Thailand; 2014 Dec 12–14. Needle in a Haystack: Reducing the Costs of Annotating Rare-Class Instances in Imbalanced Datasets; p. 244-253. 2014
20. Baldridge, J.; Osborne, M. Active Learning and the Total Cost of Annotation. In: Dekang, L.; Dekai, W., editors. *Conference on Empirical Methods in Natural Language Processing.* Barcelona, Spain: Association for Computational Linguistics; 2004. p. 9-16.
21. Snow, R.; O'Connor, B.; Jurafsky, D.; Andrew, YN. Cheap and Fast---But is it Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In: Lapata, M.; Ng, HT., editors. *Conference on Empirical Methods in Natural Language Processing.* Stroudsburg, PA: Association for Computational Linguistics; 2008. p. 254-263.
22. South, BR.; Shen, S.; Friedlin, FJ.; Samore, MH.; Meystre, SM., editors. *AMIA 2010 Symposium.* Washington DC: 2010. Enhancing Annotation of Clinical Text using Pre-Annotation of Common PHI.
23. Aberdeen J, Bayer S, Yeniterzi R, Wellner B, Clark C, Hanauer D, et al. The MITRE Identification Scrubber Toolkit: Design, training, and assessment. *Int J Med Inform.* 2010; 79(12):849–859. Epub 2010/10/19. [PubMed: 20951082]
24. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther.* 2008; 84(3):362–369. Epub 2008/05/27. [PubMed: 18500243]

25. Douglass MM, Clifford GD, Reisner A, Moore R, Marks M. Computer-assisted de-identification of free text in the MIMIC II database. *Computers in Cardiology*. 2004; 31:341–344.
26. Li R, Carrell D, Aberdeen J, Hirschman L, J K, Li B, et al. Optimizing Annotation Resources for Natural Language De-identification via a Game Theoretic Framework. *Journal of Biomedical Informatics*. 2016 (in press).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

A) Clinical note as displayed in the EHR (with PII underlined)	B) Clinical note as extracted from the EHR (with PII annotated)																																																																						
<p>[deletions]</p> <p>SUBJECTIVE: <u>Jane B Smith</u> is a pleasant <u>77</u> year old female with the above past medical history / problem list who comes in with complaint(s) of diplolia and headache. Here with her son <u>John</u></p> <p>Today she was in usual state of health. was going to a restaurant, and on way in "just hit her suddenly" with double vision.</p> <p>[deletions]</p> <p>OBJECTIVE: BP 165/82 Pulse 69 Temp 97.8 °F (36.6 °C) Resp 19 SpO2 97% Estimated Body mass index is 21.92 kg/(m^2)</p> <p>[deletions]</p> <p>LABS / STUDIES:</p> <p>HEMATOCRIT (%)</p> <table><thead><tr><th>Date</th><th>Value</th><th>Low</th><th>High</th><th>Status</th></tr></thead><tbody><tr><td><u>3/21/2011</u></td><td>46</td><td>34</td><td>49</td><td>Final</td></tr><tr><td><u>3/29/2010</u></td><td>41</td><td>34</td><td>49</td><td>Final</td></tr><tr><td><u>6/18/2010</u></td><td>39</td><td>34</td><td>49</td><td>Final</td></tr><tr><td><u>10/4/2007</u></td><td>47</td><td>34</td><td>49</td><td>Final</td></tr><tr><td><u>8/20/2008</u></td><td>41</td><td>34</td><td>49</td><td>Final</td></tr><tr><td><u>8/23/2007</u></td><td>37</td><td>34</td><td>49</td><td>Final</td></tr></tbody></table> <p>[deletions]</p> <p>ASSESSMENT / PLAN:</p> <p>1. Double vision - acute onset in patient with HTN, COPD and Hx TIA. No other symptoms of posterior circulation issues but does have acute double vision worrisome for posterior circulation. I didn't see notes of prior anisocoria but given overall mental status ? more related to cataracts. Recommend ER eval ASAP for TIA / CVA</p> <p>Spoke with <u>Jane</u>, RN at <u>Pleasant Valley Medical Center</u> ER and gave heads up</p> <p>2. HEADACHE - mild. To ER for imaging</p> <p><u>John Doe</u>, M.D. <u>(123) 456-7890</u> <u>(123) 456-7891</u> FAX</p>	Date	Value	Low	High	Status	<u>3/21/2011</u>	46	34	49	Final	<u>3/29/2010</u>	41	34	49	Final	<u>6/18/2010</u>	39	34	49	Final	<u>10/4/2007</u>	47	34	49	Final	<u>8/20/2008</u>	41	34	49	Final	<u>8/23/2007</u>	37	34	49	Final	<p>[deletions] SUBJECTIVE: <u>Jane B Smith</u> is a pleasant <u>77</u> year old female with the above past medical history / problem list who comes in with complaint(s) of diplolia and headache. Here with her son <u>John Today</u> she was in usual state of health. was going to a restaurant, and on way in "\"just hit her suddenly\"" with double vision. [deletions] OBJECTIVE: BP 165/82 Pulse 69 Temp 97.8 °F (36.6 °C) Resp 19 SpO2\r\n97% Estimated Body mass index is 21.92 kg/(m^2)</p> <p>[deletions] LABS / STUDIES: HEMATOCRIT (%)</p> <table><thead><tr><th>Date</th><th>Value</th><th>Low</th><th>High</th><th>Status</th></tr></thead><tbody><tr><td><u>3/21/2011</u></td><td>46</td><td>34</td><td>49</td><td>Final</td></tr><tr><td><u>3/29/2010</u></td><td>41</td><td>34</td><td>49</td><td>Final</td></tr><tr><td><u>6/18/2010</u></td><td>39</td><td>34</td><td>49</td><td>Final</td></tr><tr><td><u>10/4/2007</u></td><td>47</td><td>34</td><td>49</td><td>Final</td></tr><tr><td><u>8/20/2008</u></td><td>41</td><td>34</td><td>49</td><td>Final</td></tr><tr><td><u>8/23/2007</u></td><td>37</td><td>34</td><td>49</td><td>Final</td></tr></tbody></table> <p>[deletions] ASSESSMENT / PLAN:</p> <p>1. Double vision - acute onset in patient with HTN, COPD and Hx TIA. No other symptoms of posterior circulation issues but does have acute double\r\nvision worrisome for posterior circulation. I didn't see notes of prior anisocoria but given overall mental status ? more related to cataracts. Recommend ER eval ASAP for TIA / CVA Spoke with <u>Jane</u>, RN at <u>Pleasant Valley Medical Center</u> ER and gave heads up 2. HEADACHE - mild. To ER for imaging <u>John Doe</u>, M.D. <u>(123) 456-7890</u> <u>(123) 456-7891</u> FAX</p>	Date	Value	Low	High	Status	<u>3/21/2011</u>	46	34	49	Final	<u>3/29/2010</u>	41	34	49	Final	<u>6/18/2010</u>	39	34	49	Final	<u>10/4/2007</u>	47	34	49	Final	<u>8/20/2008</u>	41	34	49	Final	<u>8/23/2007</u>	37	34	49	Final
Date	Value	Low	High	Status																																																																			
<u>3/21/2011</u>	46	34	49	Final																																																																			
<u>3/29/2010</u>	41	34	49	Final																																																																			
<u>6/18/2010</u>	39	34	49	Final																																																																			
<u>10/4/2007</u>	47	34	49	Final																																																																			
<u>8/20/2008</u>	41	34	49	Final																																																																			
<u>8/23/2007</u>	37	34	49	Final																																																																			
Date	Value	Low	High	Status																																																																			
<u>3/21/2011</u>	46	34	49	Final																																																																			
<u>3/29/2010</u>	41	34	49	Final																																																																			
<u>6/18/2010</u>	39	34	49	Final																																																																			
<u>10/4/2007</u>	47	34	49	Final																																																																			
<u>8/20/2008</u>	41	34	49	Final																																																																			
<u>8/23/2007</u>	37	34	49	Final																																																																			
<p>Key: <u>Annotation correct</u> <u>Annotation too long</u> <u>Annotation too short</u> <u>Annotation missing</u></p> <p>[deletions] = Where clinical note text was deleted to shorten note for presentation here</p> <p>This clinical note has been fictionalized to preserve patient privacy.</p>																																																																							

Figure 1.

Comparison of a clinical note (A) as displayed in the EHR with actual PII underlined, and (B) as extracted from the EHR with color highlighting representing annotations that are correct (green), too long (blue) too short (yellow) or missing (underlined without highlighting).

Figure 2.

Median number of newly discovered PII instances discovered by annotator teams of increasing size (individuals, simulated teams of 2, 3, or 4) and salary cost per additional PII instance discovered in 2014 dollars.

Table 1

Summary of personally identifiable information (PII) by PII type in a corpus of 100 * Family Practice clinical notes.

PII type	HIPAA?	Count	Percent
Date (or part, including year)	Yes	569	52.1%
Organization name (e.g., hospital)	No	158	14.5%
Practitioner name	No	122	11.2%
Patient age	Yes	90	8.2%
Patient name	Yes	78	7.1%
Other identifier	Yes	20	1.8%
Practitioner initials	No	19	1.7%
Geographic locations	Yes	14	1.3%
Telephone number	Yes	13	1.2%
Medical record number	Yes	5	0.5%
Web site	Yes	5	0.5%
All PII types		1,093	100%

* Six of the 100 notes did not contain PII.

Summary of instances of personally identifiable information (PII) contributed to the gold standard (GS) by all four, only 3, only 2, or only 1 annotator(s).

Table 2

N annotators annotating each GS PII instance	Row	GS PII was annotated by annotator(s)				PII instances annotated only by annotators marked "Yes" in row or group (a)			
		1	2	3	4	Row count	Row percent (b)	Group count	Group percent (b)
All 4 annotators	4-A	Yes	Yes	Yes	Yes	915	83.7%	915	83.7%
Only 3 annotators	3-A	Yes	Yes	Yes	No	95	8.7%		
	3-B	Yes	Yes	No	Yes	13	1.2%		
	3-C	Yes	No	Yes	Yes	24	2.2%	155	14.2%
	3-D	No	Yes	Yes	Yes	23	2.1%		
Only 2 annotators	2-A	Yes	Yes	No	No	4	0.4%		
	2-B	Yes	No	Yes	No	3	0.3%		
	2-C	Yes	No	No	Yes	0	0.0%		
	2-D	No	Yes	Yes	No	2	0.2%	17	1.6%
	2-E	No	Yes	No	Yes	3	0.3%		
	2-F	No	No	Yes	Yes	5	0.5%		
Only 1 annotator	1-A	Yes	No	No	No	0	0.0%		
	1-B	No	Yes	No	No	2	0.2%		
	1-C	No	No	Yes	No	2	0.2%	6	0.5%
	1-D	No	No	No	Yes	2	0.2%		
Annotator PII total:		1054	1057	1069	985	1093	100.0%	1093	100.0%
Annotator recall (c):		96.4%	96.7%	97.8%	90.1%				

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Notes:

- (a) Rows and groups of rows indicate individual annotators or combinations of annotators making unique PII contributions to the gold standard (i.e., PII annotated by that individual or group and none other). For example, row 2-B indicates that annotators 1 and 3 identified 3 instances of PII not identified by any other individual or group.
- (b) Row percent and group percent are calculated as the number of PII instances annotated by annotators marked "Yes" in that row or group divided by 1,093, which is the total number of PII instances in the gold standard.
- (c) Annotator recall is calculated as the number of PII instances annotated by the individual annotator divided by 1,093, which is the total number of PII instances in the gold standard.

Table 3

Number of personally identifiable information (PII) instances annotated, PII instances overlooked, and PII recall rate for teams of one, two, and three annotators with summary statistics by annotator team size.

Annotator team size	Annotator(s)*	PII instances annotated	PII instances overlooked	PII recall rate
One annotator	A1	1,054	39	0.964
	A2	1,057	36	0.967
	A3	1,069	24	0.978
	A4	985	108	0.901
	Median:	1,056	38	0.97
	Mean:	1,041.3	51.75	0.953
	Std. dev.:	38.06	38.06	0.035
Two annotators	A1 + A2	1,084	9	0.992
	A1 + A3	1,086	7	0.994
	A1 + A4	1,087	6	0.995
	A2 + A3	1,091	2	0.998
	A2 + A4	1,088	5	0.995
	A3 + A4	1,087	6	0.995
	Median:	1,087	6	0.99
Three annotators	A1 + A2 + A3	1,091	2	0.998
	A1 + A2 + A4	1,091	2	0.998
	A1 + A3 + A4	1,091	2	0.998
	A2 + A3 + A4	1,093	0	1.000
	Median:	1,091	2	1.00
	Mean:	1,091.5	1.50	0.999
	Std. dev.:	1.00	1.00	0.001

* A1, A2, A3, and A4 = annotators #1, #2, #3, & #4, respectively.

Table 4

Recall rates for annotation of personally identifiable information (PII) by teams of one, two, and three annotators by PII type.

PII Type and Annotation Team Size	Recall Rate			
	Min	Median	Max	Std. Dev.
<i>Date</i>				
<i>One annotator</i>	0.963	0.983	0.989	0.012
<i>Two annotators</i>	0.995	0.998	1.000	0.002
<i>Three annotators</i>	0.998	1.000	1.000	0.001
<i>Organization Name</i>				
<i>One annotator</i>	0.918	0.949	0.994	0.033
<i>Two annotators</i>	0.981	1.000	1.000	0.008
<i>Three annotators</i>	1.000	1.000	1.000	0.000
<i>Practitioner Name</i>				
<i>One annotator</i>	0.967	0.971	0.992	0.012
<i>Two annotators</i>	0.975	0.988	1.000	0.010
<i>Three annotators</i>	0.984	1.000	1.000	0.008
<i>Patient Age</i>				
<i>One annotator</i>	0.000	0.972	1.000	0.736
<i>Two annotators</i>	0.956	0.994	1.000	0.017
<i>Three annotators</i>	0.989	1.000	1.000	0.006
<i>Patient Name</i>				
<i>One annotator</i>	0.974	0.994	1.000	0.012

PII Type and Annotation Team Size	Recall Rate				Std. Dev.
	Min	Median	Max	Mean	
<i>Two annotators</i>	0.987	1.000	1.000	0.998	0.005
<i>Three annotators</i>	1.000	1.000	1.000	1.000	0.000
<i>Other identifier</i>					
<i>One annotator</i>	0.950	1.000	1.000	0.988	0.025
<i>Two annotators</i>	1.000	1.000	1.000	1.000	0.000
<i>Three annotators</i>	1.000	1.000	1.000	1.000	0.000
<i>Practitioner initials</i>					
<i>One annotator</i>	0.895	0.895	0.895	0.895	0.000
<i>Two annotators</i>	0.895	0.947	1.000	0.947	0.058
<i>Three annotators</i>	0.895	1.000	1.000	0.974	0.053
<i>Geographical Locale</i>					
<i>One annotator</i>	0.786	0.929	1.000	0.911	0.090
<i>Two annotators</i>	0.929	1.000	1.000	0.988	0.029
<i>Three annotators</i>	1.000	1.000	1.000	1.000	0.000
<i>Phone Number</i>					
<i>One annotator</i>	0.692	0.923	1.000	0.885	0.147
<i>Two annotators</i>	0.846	1.000	1.000	0.974	0.063
<i>Three annotators</i>	1.000	1.000	1.000	1.000	0.000
<i>Medical Record Number</i>					
<i>One annotator</i>	1.000	1.000	1.000	1.000	0.000

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

PII Type and Annotation Team Size	Recall Rate				
	Min	Median	Max	Mean	Std. Dev.
Two annotators	1.000	1.000	1.000	1.000	0.000
Three annotators	1.000	1.000	1.000	1.000	0.000
Web/URL/email					
One annotator	0.800	1.000	1.000	0.950	0.100
Two annotators	1.000	1.000	1.000	1.000	0.000
Three annotators	1.000	1.000	1.000	1.000	0.000