Journal of Biomedical Informatics 50 (2014) 173–183

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research

Louise Deleger, Todd Lingren [1], Yizhao Ni [1], Megan Kaiser, Laura Stoutenborough, Keith Marsolo, Michal Kouril, Katalin Molnar, Imre Solti *

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

ARTICLE INFO

Article history:
Received 1 September 2013
Accepted 30 January 2014
Available online 17 February 2014

Keywords:
Natural Language Processing
Privacy of patient data
Health insurance portability and accountability act
Automated de-identification
De-identification gold standard
Protected Health Information

ABSTRACT

Objective: The current study aims to fill the gap in available healthcare de-identification resources by creating a new sharable dataset with realistic Protected Health Information (PHI) without reducing the value of the data for de-identification research. By releasing the annotated gold standard corpus with Data Use Agreement we would like to encourage other Computational Linguists to experiment with our data and develop new machine learning models for de-identification. This paper describes: (1) the modifications required by the Institutional Review Board before sharing the de-identification gold standard corpus; (2) our efforts to keep the PHI as realistic as possible; (3) and the tests to show the effectiveness of these efforts in preserving the value of the modified data set for machine learning model development.

Materials and methods: In a previous study we built an original de-identification gold standard corpus annotated with true Protected Health Information (PHI) from 3503 randomly selected clinical notes for the 22 most frequent clinical note types of our institution. In the current study we modified the original gold standard corpus to make it suitable for external sharing by replacing HIPAA-specified PHI with newly generated realistic PHI. Finally, we evaluated the research value of this new dataset by comparing the performance of an existing published in-house de-identification system, when trained on the new de-identification gold standard corpus, with the performance of the same system, when trained on the original corpus. We assessed the potential benefits of using the new de-identification gold standard corpus to identify PHI in the i2b2 and PhysioNet datasets that were released by other groups for de-identification research. We also measured the effectiveness of the i2b2 and PhysioNet de-identification gold standard corpora in identifying PHI in our original clinical notes.

Results: Performance of the de-identification system using the new gold standard corpus as a training set was very close to training on the original corpus (92.56 vs. 93.48 overall *F*-measures). Best i2b2/PhysioNet/CCHMC cross-training performances were obtained when training on the new shared CCHMC gold standard corpus, although performances were still lower than corpus-specific trainings.

Discussion and conclusion: We successfully modified a de-identification dataset for external sharing while preserving the de-identification research value of the modified gold standard corpus with limited drop in machine learning de-identification performance.

1. Introduction

The current study aims to fill the gap in available healthcare de-identification resources by creating a new sharable dataset with realistic Protected Health Information (PHI) without reducing the value of the data for de-identification research. By releasing the annotated gold standard corpus with Data Use Agreement we would like to encourage other Computational Linguists to experiment with our data and develop new machine learning models for de-identification. This paper describes: (1) the modifications required by the Institutional Review Board before sharing the de-identification gold standard corpus; (2) our efforts to keep the PHI as realistic as possible; (3) and the tests to show the effectiveness of these efforts in preserving the value of the modified data set for machine learning model development.

Abbreviations: NLP, Natural Language Processing; PHI, Protected Health Information; DUA, Data Use Agreement; IRB, Institutional Review Board; ML, machine learning; CCHMC, Cincinnati Children's Hospital Medical Center.

* Corresponding author. Address: Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MLC 7024, Cincinnati, OH 45229-3039, USA. Fax: +1 513 636 2056.

E-mail address: imre.solti@cchmc.org (I. Solti).

[1] Equal contribution.

http://dx.doi.org/10.1016/j.jbi.2014.01.014
1532-0464/

The new resource consists of over 3500 notes, 22 clinical note types, and includes all HIPAA-specified PHI classes. The data set is available for de-identification research immediately. Interested parties should contact the senior author.

The motivation of this effort stems from lack of sharable de-identification datasets. We will describe: (1) the modifications necessary for the original corpus to achieve Institutional Review Board (IRB) and legal approval of the data release with a Data Use Agreement (DUA); (2) the simultaneous efforts to preserve the de-identification research value of the original data; (3) the approaches to minimize the use of synthetic (i.e. "fake") PHI while balancing IRB and legal constraints; and (4) the evaluation methodology to compare the new and the original datasets' de-identification research value.

Gold standard annotated corpora are necessary resources when building and evaluating Natural Language Processing (NLP) systems. Manually labeled instances that are relevant to the specific NLP tasks must be created. A useful gold standard should be rich in information and include large variety of documents and annotated instances that represent the diversity of document types and instances at stake in a specific task. This is essential to (1) either train machine-learning based NLP systems, which need examples to learn from, or discover rules for rule-based algorithms and (2) evaluate the performance of NLP systems.

De-identification of clinical narrative text is usually a necessary preliminary step for many research tasks that include sharing the data with researchers outside of the healthcare entity that generated the data. De-identification systems that remove PHI are examples of systems that require carefully annotated gold standard corpora.

Automated NLP-based de-identification methods have been developed [1] and a number of non-shared corpora have been built for measuring their performance. These corpora present some limitations as they often consist of only a few document types, such as discharge summaries [2,3], nursing notes, pathology reports [4–6], outpatient follow-up notes [7], or medical message boards [8]. They do not always include annotation of all required PHI, e.g. locations and contact information are sometimes ignored [6,9].

Currently, only two de-identification corpora are available to the public with DUA: (1) the corpus from the i2b2 2006 de-identification shared task [3] and (2) the PhysioNet corpus [10]. The i2b2 corpus is a collection of discharge summaries, while PhysioNet's consists of nursing notes. Both corpora contain synthetic PHI, i.e. they have been de-identified and true PHI have been replaced with surrogate PHI. True PHI is defined by the PHI which are present in the original text. Surrogate PHI are substituted PHI from one or more sources to replace the true PHI. The i2b2 corpus has a total of 889 documents, including 19,498 PHI, while the PhysioNet corpus contains 2483 documents but with a very limited density of PHI (1779).

Our work aims to extend existing sharable resources and overcome some of their limitations, by proposing a de-identified corpus with a larger diversity of note types (over 22 different note types), and a larger set of documents (over 3500 notes) and PHI annotations (over 30,000 annotations).

There are variations in resynthesis processes used to replace PHI in corpora [3,11,12]. For PHI involving numerical values such as dates, ids and phone numbers, approaches are usually based on digit replacement strategies. Approaches for names are less similar. Uzuner et al. [3] focused on generating a majority of out-of-vocabulary names, while Yeniterzi et al. [11] used names from a dictionary. In their work on Swedish clinical notes, Alfalahi et al. [12] used names from dictionaries while also introducing some letter variations to allow for misspelled names, and kept the gender intact in first names. Our approach draws on those methods, while also introducing some novel replacement strategies.

Using synthetic PHI to train for de-identification might introduce bias for de-identifying real data. Yeniterzi et al. examined the effect and potential bias that a corpus with surrogate PHI can have on clinical text de-identification [11]. They built a corpus (not shared externally) composed of four classes of clinical narrative texts (laboratory reports, medication orders, discharge summaries and physician letters) and replaced the original PHI with synthetic PHI using the resynthesis engine of the MITRE Identification Scrubber Toolkit [13]. They showed that machine-learning-based de-identification achieved high performance when using homogeneous training and test sets (either the original corpus (F-measure = 0.96) or the re-synthetized corpus (0.98)), while performance declined significantly when training on the resynthetized corpus to de-identify real data (0.728).

Besides advancing research on de-identification of clinical text, de-identified clinical corpora can also be useful for research on clinical NLP in general, for instance developing clinical information extraction systems. Because of privacy issues, clinical corpora are not widely available for research purposes or to research teams outside of healthcare institutions. De-identified clinical corpora which can be shared with DUAs can therefore help advance research on clinical NLP, which is a secondary motivation behind our work.

2. Materials and methods

We distinguish two main steps in building our annotated corpus for external sharing. First, we created the original de-identification corpus annotated with true PHI. Second, we modified that corpus to make it suitable for external sharing.

This study was conducted under an approved Institutional Review Board (IRB) protocol.

2.1. Original de-identification corpus

We provide only a summary of step one. Details of that effort are described in two earlier publications [14,15]. Our original corpus is composed of 3503 clinical notes selected by stratified random sampling from five million notes composed by Cincinnati Children's Hospital Medical Center (CCHMC) clinicians during 2010. The notes include over 22 different note types (Table 1). We included a variety of note types (discharge summaries, ED notes, etc.). We selected a note type only if the number of notes exceeded the subjective limit of 800 during the previous 12-month period. We oversampled discharge summaries because of their richness in deidentification information.[13] We also oversampled some of the notes that were less frequent but exceeded the 800-note limit to have at least 20 notes for each type in the study set. The total number of note types in the final study set is above 22. Because of the way our EHR is configured, some of the note types have no labels (e.g., "external notes" contain diagnostic test reports such as radiology reports, but have no specific labels). More details can be found in prior publications [14,15].

The clinical notes were double annotated for PHI by two annotators. We defined 12 classes of PHI, derived from and extending the 18 HIPAA categories:

- **NAME**: any first name, middle name, last name or combination of those.
- **DATE**: date (e.g. "12/29/2005").
- **AGE**: age of the patient (any age).
- **EMAIL**.
- **INITIALS**: initials of a person (occurring on their own).

Table 1
List of note types in the corpus.

Type	Number of notes
Asthma Action Plan	40
Brief OpNote	40
Communication Body	40
Consult Note	40
DC Summaries	400
ED Medical Student	40
ED Notes	218
ED Provider Notes	111
ED Provider Reassess.	24
H&P	20
Med Student	20
Operative Report	20
OR Nursing	20
Patient Instructions	33
Pharmacy Note	20
Plan of Care Note	75
Pre-Op_Evaluation	20
Procedure Note	20
Progress Notes Outp	179
Progress Notes Inp	128
Referral	20
Telephone Encounter	127
All labeled notes	**1655**
Unlabeled notes	**649**
External notes	**1199**
All notes	**3503**

- **INSTITUTION**: hospital names and other organizations.
- **IPADDRESS**: includes IP addresses and URLs.
- **LOCATION**: geographical locations such as address and city.
- **PHONE**: phone and fax numbers.
- **SSN**: social security number.
- **IDNUML**: any identification number such as medical record number, and patient ID.
- **OTHER**: internal locations inside a hospital (e.g. ER)

In theory, the "OTHER" category also includes all other potential protected information (such as medical device serial numbers and license plate numbers). However clinical notes from our corpus do not contain any such information, so the "OTHER" category only contains internal locations.

Inter-annotator agreement and other relevant corpus statistics are presented in the two earlier publications [14,15].

2.2. Building the new de-identification gold standard corpus

The motivation for modifying the annotated gold standard corpus before sharing it arises from privacy and legal requirements to protect patients' rights and try to eliminate legal liability for the institution releasing the data. From the original manually-annotated CCHMC corpus, we created a new version that is suitable for external sharing by replacing HIPAA-specified PHI with newly generated PHI. At the same time, it is equally important that the modified dataset retains its de-identification research value, for example, for machine-learning training purposes. Consequently, the modifications should be minimally intrusive for research purposes, while completely satisfying privacy and legal requirements.

As not all PHI annotated in our corpus are required to be de-identified by HIPAA, a number of PHI classes were kept unchanged. This was the case for INSTITUTION, INITIALS, OTHER and AGE (HIPAA only requires ages > 89 to be de-identified, which our corpus does not contain).

All other PHI classes were replaced, using methods described below and exemplified in Fig. 1. Special care was given to insure that (1) the real PHI could not be determined by studying its surrogate, (2) most replacements were linguistically coherent with the original PHI and the context in which they occurred and (3) multiple instances of an element corresponded to identical replacement PHI. For instance, we tried to keep the discourse coherent by replacing elements of the PHI with new elements following the same pattern (e.g. we replaced date occurrences such as "November" with similar elements, for example "January", but not with elements such as "11/08/2013"), and also by keeping the gender when replacing names. The various external resources used to replace PHI were gathered by the MIST de-identification software developers and were provided with the MIST toolkit [13].

DATE phrases were grouped into several categories, each corresponding to a specific pattern, and shuffled throughout the corpus (i.e., each DATE phrase was replaced by a different DATE phrase from the corpus). Patterns described the various possible date formats such as "11/19/2011", "November 19th 2011" and "11–19". For linguistic coherence, each DATE PHI was replaced with a different PHI with a similar format (e.g. "January 17th" was replaced with "November 5th" and "12/05/2008" was replaced with "06/12/2005" in Fig. 1). Days of the week that were part of a DATE element (e.g., "Monday, January 10") were included in the replacement process, per requirement of our legal department. As it is difficult to have an exhaustive categorization of all possible patterns, we also created a special category for dates which did not correspond to any specified pattern. None of the replaced PHIs within the same clinical note occurred in the original version of the note. That is only DATE phrases from different documents were used to replace a DATE PHI in a given document.

EMAIL phrases were replaced with fake emails, composed of randomly generated sequences of letters of the same length as the original sequences. For instance, john.smith@cchmc.org was replaced with rgnv.fgbtuk@joitq.btg in Fig. 1. All generated PHI were different from all original EMAIL PHIs occurring (anywhere) in the original corpus. Based on the request of the IRB we chose to generate non-realistic email addresses, to avoid creating addresses that might actually exist. But this also means that for human fake emails are easily distinguishable from real emails. This might be a problem if emails have been missed in the de-identification process, and this is a limitation of our replacement approach. Since our corpus has been manually double-annotated, we do not expect this phenomenon to occur in our corpus.

IPADDRESS phrases were replaced by URLs randomly selected from a list of URLs (available in the MIST package [13]). All generated IPADDRESS phrases were different from all IPADDRESS phrases occurring (anywhere) in the original corpus.

IDNUM phrases were replaced with randomly generated sequence of numbers of the same length as the original PHI. For example, medical record number "214337" was replaced with "439251" in Fig. 1. All generated IDNUM phrases were strictly different from all original IDNUMs occurring (anywhere) in the original corpus. The number zero was not allowed in the first digit position for the IDNUM phrases because none of the original IDNUMs had zero in the first position either.

Area codes from *PHONE phrases* were replaced by new PHONE phrases randomly selected from a list of all US area codes (available in the MIST package [13]). All other sequences of numbers in PHONE PHIs (including international country codes, if any) were replaced by randomly generated sequences of numbers, of the same length as the original PHONE phrases (the number zero was not allowed in the first digit position for the first 3 digits after the area code). For instance, "513-659-8995" was replaced by "201-523-6611" in Fig. 1. All generated PHONE phrases were different from all original PHONE PHIs occurring (anywhere) in the original corpus.

NAME PHIs were replaced by new NAME phrases with the same pattern. That is, the names were parsed to recognize their various

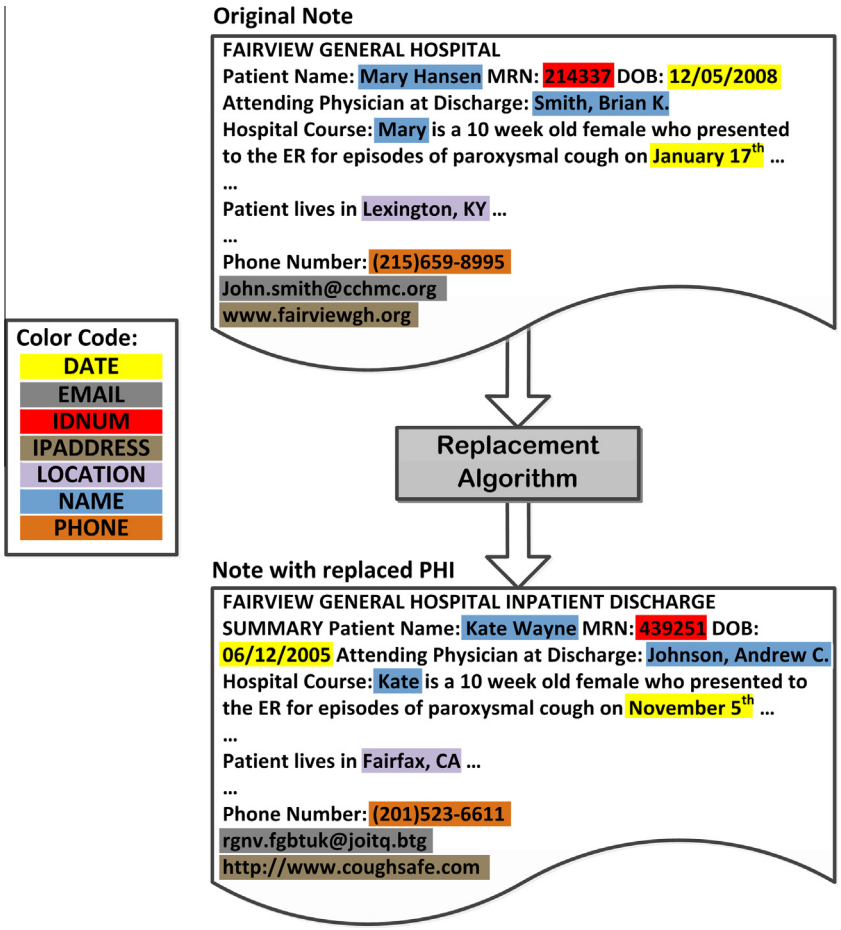

Fig. 1. Example of PHI replacement in a clinical note.

possible tokens. Name patterns included: first name alone, last name alone, first name + last name, first name + initial + last name, first name + middle name + last name, etc. Each token of a NAME phrase was then replaced by a new token selected from lists of names provided in the MIST package [13] and originating from the US Census Bureau [16]. Lists include a list of female first names, a list of male first names and a list of last names. Based on the request of the IRB, only the first names and last names that occurred in at least 0.002% (absolute frequency of 144) of the US Census Bureau's data were used to generate new names. The above process resulted in a list of 11,764 names in total, with 7499 last names, 3050 female first names and 1215 male first names.

Original first names and middle names from our corpus were replaced by first names randomly selected from the lists. Female names were replaced by female names, male names by male names, and first names that could be either female or male by first names that could be either female or male. When the original first name was not present in the lists, a random name was picked and was altered by replacing two of its letters with random characters, so that the generated name would be different from all listed first names. Last names were replaced by last names randomly selected from the last name list. A similar process was followed for last names that did not appear in the list of last names (by introducing letter alteration). Our replacement strategy shares similarities with the approach of Alfalahi et al. [12] conducted for Swedish name replacement, as their method was also designed to take into account out-of-vocabulary/misspelled names and gender. Initials were replaced by randomly selected uppercase letters.

As much as possible, continuity was preserved in the new corpus and the same first names were replaced by the same replacement first names, and the same last names by the same replacement last names. An example of name replacements is shown in Fig. 1 where full patient name "Mary Hansen" was replaced with "Kate Wayne" (i.e., first name "Mary" was replaced with "Kate" and last name "Hansen" was replaced with "Wayne"), then the patient first name occurring on its own was replaced with "Kate" (keeping the same first name as in the full name) and physician name "Smith, Brian K." was replaced with "Johnson, Andrew C.". To insure that generated NAME phrases could not give any indication to the real names, the following precautions were taken:

- Throughout the entire corpus, all generated last names were different from all original last names. An original last name was never used to create a new name if it appeared anywhere in the corpus.
- Within a clinical note, all tokens of all new NAME phrases were different from all the original tokens of all the original NAME phrases in that note. That is, in addition to the above (corpus-level) restriction, an original first name was never used to create a new name if it appeared anywhere in the original note.

LOCATION phrases were parsed to recognize the various possible tokens of a location and replaced by phrases following the same pattern. Patterns included state name, city name + state name, address (e.g. street number + street name + street suffix + city name + zip code + state), institution name + address, etc. Street

names, city names, zip codes, and state names were shuffled throughout the corpus, and each token was replaced by a random (and different) LOCATION token from the corpus. Street numbers were replaced by random sequences of numbers of the same length. When re-inserting the shuffled LOCATION tokens in the new corpus, we never paired elements that were occurring together in the original corpus. Newly generated pairs of street names and street suffixes were always different from the LOCATION tokens appearing in the original corpus (e.g. if "Martin Luther King" and "drive" were seen together anywhere in the original corpus, then they were never paired together in the new corpus), and the same rule was applied for pairs of zip codes, city names, and state names. Consequently, an address in the new corpus is different from all addresses of the original corpus. Institution names that were part of an address (e.g. "CCHMC, 3333 Burnet Avenue") were also shuffled through the corpus, and each institution name was replaced by a random (and different) token from the corpus. In Fig. 1, the location "Lexington, KY" was replaced with the location "Fairfax, CA", which has the same pattern.

The single occurrence of a *social security number (SSN)* in the original corpus was replaced by "xx-xxx-xxxx" pattern.

2.3. Experiments

2.3.1. Descriptive statistics

We computed PHI descriptive statistics for the two corpora (the original corpus and the new corpus) and compared the frequency distributions in terms of unique PHIs. We assessed if there was a frequency imbalance between the two corpora. A large frequency imbalance could have an impact on the performance of machine learning systems, if they are trained on a corpus with little variability and applied to a corpus with more variability. A machine-learning system trained on a corpus with many repeated PHI phrases is more likely to give an important weight to the tokens themselves as opposed to other crucial features such as the context of occurrence [17]. Consequently, the system will miss many PHI phrases when confronted to a corpus with more variety.

2.3.2. Machine learning de-identification performance on original CCHMC notes

We investigated the usefulness of the newly created corpus in de-identifying real data by comparing the performance of an unmodified existing in-house de-identification system, trained on the new corpus, with the performance of that same system trained on the original corpus [14,15]. The motivation was to evaluate how well the corpus retained its de-identification research value after the privacy driven modifications. Performance was tested against the original corpus, in a ten-fold cross-validation setting. In each cross-validation set, 90% of the corpus (3153 documents) was used to train the system while 10% (350 documents) was used for testing.

The in-house system used for the experiments is a hybrid NLP system that combines linguistic pre-processing and rule-based post-processing with a core machine-learning algorithm (Conditional Random Fields as implemented in the MALLET toolkit [18]). It is described in detail in an earlier publication [14].

2.3.3. Machine learning de-identification performance on CCHMC, i2b2 and PhysioNet corpora through cross training

In addition to the original CCHMC notes, we also assessed the potential benefit of using our new sharable corpus to de-identify other available resources that were released by other research groups. We also measured the effectiveness of the other sharable resources to train our machine-learning algorithm to de-identify the original CCHMC dataset. Our working hypothesis was that, while the less PHI-rich i2b2 and PhysioNet datasets would not improve the de-identification of more diverse and PHI-rich CCHMC notes statistically significantly, the newly released corpus would improve the de-identification performance of the existing datasets. Three main experiments were conducted.

First we tested our de-identification system on **the i2b2 2006 dataset** [3], which consisted of de-identified discharge summaries (669 reports for training and 220 reports for testing). We kept the original distribution between training and test sets, as described in the i2b2 challenge. We trained our system using:

1. The CCHMC new corpus, alone and in combination with 10% increments of the i2b2 dataset.
2. The PhysioNet corpus, alone and in combination with increasingly larger percentages of the i2b2 dataset.
3. The i2b2 corpus alone (similarly, in sets of increasing size, for comparison purposes).

Since the i2b2 dataset has slightly different PHI classes than our corpus and the PhysioNet corpus, we had to adjust PHI classes to harmonize classes across the three corpora. We removed the EMAIL, IPADDRESS, INITIALS, and OTHER classes from the CCHMC corpus. We merged the DOCTOR and PATIENT classes as NAME in the i2b2 dataset, and the health care provider name, patient name and relative name as NAME in the PhysioNet corpus. The above steps resulted in seven classes to train and evaluate the system: AGE, DATE, IDNUM, INSTITUTION, LOCATION, NAME, PHONE.

We performed the same experiment **on the PhysioNet dataset** [2,10]. We tested the performance of the system on this corpus, in a ten-fold cross-validation setting when training on:

1. The CCHMC new corpus, alone and in combination with 10% increments of the PhysioNet corpus.
2. The i2b2 corpus, alone and in combination with the PhysioNet corpus.
3. The PhysioNet corpus alone.

Similarly to the i2b2 experiment, we also adjusted the PHI classes. We removed the EMAIL, IPADDRESS, INSTITUTION, and OTHER classes from the CCHMC corpus. We removed the HOSPITAL class from the i2b2 dataset and we merged the DOCTOR and PATIENT classes as NAME. We merged the health care provider name, patient name and relative name as NAME in the PhysioNet corpus. The above steps resulted in seven classes to train and evaluate the system: AGE, DATE, IDNUM, INITIALS, LOCATION, NAME, PHONE.

The third main experiment assessed the value of the i2b2 and PhysioNet datasets to de-identify **the CCHMC original corpus**. We tested the performance of the system on this corpus, in a ten-fold cross-validation setting, when training on:

1. the i2b2 corpus, alone and in combination with 10% increments of the CCHMC original corpus;
2. the PhysioNet corpus, alone and in combination with the CCHMC original corpus;
3. the CCHMC original corpus alone.

For this experiment, we used the 12 original classes of our CCHMC corpus (AGE, DATE, EMAIL, IDNUM, INITIALS, INSTITUTION, IPADDRESS, LOCATION, NAME, OTHER, PHONE, SSN) to evaluate the system. We adjusted name classes in the i2b2 and PhysioNet dataset similarly as above, by merging all different name classes as a single NAME class.

Table 2 gives a summary of the three cross-corpus experiments, showing their main characteristics, the various training configurations used, and in which table/figure results will be presented.

2.4. Metrics

We tested statistical significance of the difference in unique PHI frequency, in the original and the new CCHMC corpora, with Pearson's Chi-square.

Performance of the de-identification system was measured using standard metrics in NLP, which are precision (P), recall (R) and F-measure (F) [19,20]. F-measure is the weighted harmonic mean of precision and recall and can be calculated as ($\beta = 1$): $F = (1 + \beta^2)(P^*R)/\ \beta^{2*}(P + R)$. We computed performance measurements for each individual PHI category and overall (at the exact span level) in ten-fold cross validations.

To rule out the possibility that the performance difference between two outputs was due to chance, we also tested the statistical significance of the difference, using a computationally intensive method known as approximate randomization [21,22] which is not dependent on the distribution of the underlying data. Due to the number of different significance tests conducted, we applied a Bonferroni correction to account for the increased possibility of Type I error [23]. A type I error occurs when one rejects the null hypothesis when it is actually true. When multiple tests are performed on a single dataset, the probability of incorrectly rejecting the null hypothesis increases as more hypotheses are tested.

3. Results

3.1. Descriptive statistics

Table 3 shows the descriptive statistics of the original and the new CCHMC corpora. Table 4 shows similar statistics for the i2b2 and PhysioNet data sets. Differences in the distribution of unique PHI in the two CCHMC corpora are never statistically significant for any PHI category (last column of Table 3). The number of PHI in the CCHMC corpus is larger than that of either the i2b2 or the PhysioNet corpus, in total and per PHI category, with the exception of INSTITUTION and IDNUM, which are more numerous in the i2b2 corpus. The CCHMC corpus is also the largest corpus (over one million tokens).

3.2. De-identification performance

3.2.1. Machine learning de-identification performance on original CCHMC notes

Performance of the automated de-identification system using the new corpus as training set was slightly lower than, but very close to the performance when training on the original corpus for each individual PHI category and overall (0.9256 vs. 0.9348 overall F-measures (Table 5)). The difference was statistically significant for the NAME and IDNUM categories and overall. To adjust for the 13 different significance tests conducted, the performance was considered statistically significant at p-values < 0.0038 (i.e. 0.05/13 using Bonferroni correction).

3.2.2. Machine learning de-identification performance on CCHMC, i2b2 and PhysioNet corpora through cross-training

Table 6 presents performance (overall precision, recall and F-measure) obtained in the three main cross-training experiments, when training only on one corpus and testing on another, along with performance obtained with same-corpus training (for comparison purposes). Figs. 2–4 visualize performance when combining corpora with increasing percentages of the dataset on which the evaluation is focused (i2b2, PhysioNet and original CCHMC, respectively).

3.2.2.1. Performance on the i2b2 corpus. When testing on the i2b2 corpus but training only with the new CCHMC corpus or the PhysioNet corpus, performance is low (overall F-measures of 0.4705 and 0.1558) compared to training on the same source corpus (the i2b2 corpus). However, compared with the PhysioNet experiment, the F-measure was improved by 200% in the CCHMC experiment (Table 6). Fig. 2 displays the performance when the new CCHMC corpus and the PhysioNet corpus are combined with increasing percentages of the i2b2 corpus. It shows that combining the new CCHMC corpus and the i2b2 corpus yielded almost always higher performance than when using the i2b2 corpus alone, although the difference is limited, as is demonstrated by the close proximity of the i2b2 and i2b2 + new CCHMC dots in the graphs. Performance when combining the PhysioNet corpus with the

Table 2
Summary of cross-training experiments.

Training corpus	Test corpus	PHI classes	Results
i2b2 experiment			
i2b2	i2b2	AGE, DATE, IDNUM, INSTITUTION, LOCATION, NAME, PHONE	Table 6
new CCHMC			
PhysioNet			
10% increments of the i2b2 corpus			Fig. 2
New CCHMC + 10% increments of the i2b2 corpus			
PhysioNet + 10% increments of the i2b2 corpus			
PhysioNet experiment			
PhysioNet	PhysioNet	AGE, DATE, IDNUM, INITIALS, LOCATION, NAME, PHONE	Table 6
new CCHMC			
i2b2			
10% increments of the PhysioNet corpus			Fig. 3
New CCHMC + 10% increments of the PhysioNet corpus			
i2b2 + 10% increments of the PhysioNet corpus			
Original CCHMC experiment			
Original CCHMC	Original CCHMC	AGE, DATE, EMAIL, IDNUM, INITIALS, INSTITUTION, IPADDRESS, LOCATION, NAME, OTHER, PHONE, SSN	Table 6
i2b2			
PhysioNet			
10% increments of the original CCHMC corpus			Fig. 4
i2b2 + 10% increments of the original CCHMC corpus			
PhysioNet + 10% increments of the original CCHMC corpus			

Table 3
Number of PHI in the original and the new CCHMC corpora.

PHI category	Frequency in original CCHMC (token count = 1,072,957)		Frequency in new CCHMC (token count = 1,076,198)		Statistical significance (*p*-value)
	Total	Unique	Total	Unique	
AGE	2109	776	2109	776	1
DATE	13,060	3283	13,060	3154	0.064
EMAIL	14	12	14	12	1
IDNUM	1117	1062	1117	1062	1
INITIALS	16	12	16	12	1
IPADDRESS	10	5	10	5	1
INSTITUTION	1994	389	1994	389	1
LOCATION	396	270	396	244	0.053
NAME	7776	4024	7776	4042	0.773
OTHER	3446	685	3446	685	1
PHONE	876	480	876	480	1
SSN	1	1	1	1	1
Total	30,815	10,999	30,815	10,862	0.106

Table 4
Number of PHI in the i2b2 and PhysioNet corpora.

PHI category	Frequency in i2b2 corpus (token count = 554,999)		Frequency in PhysioNet corpus (token count = 335,383)	
	Total	Unique	Total	Unique
AGE	16	14	4	1
DATE	7098	1363	528	285
DOCTOR NAME	3751	2587	593	410
INITIALS	–	–	2	2
INSTITUTION	2400	646	–	–
IDNUM	4809	4431	–	–
LOCATION	263	244	367	143
OTHER	–	–	3	3
PATIENT NAME	929	596	54	39
PHONE	232	141	53	45
RELATIVE NAME	_	_	175	121
Total	19,498	10,022	1779	1049

Table 5
De-identification system performance when training on the original vs on the new corpus.

	Training on original			Training on new			Statistical significance (*p*-value)
	P	*R*	*F*	*P*	*R*	*F*	
AGE	0.9669	0.9	0.9322	0.9623	0.8962	0.9281	0.053
DATE	0.9795	0.9698	0.9746	0.9822	0.9617	0.9719	0.0489
EMAIL	0.9333	1	0.9655	0.9333	1	0.9655	1
IDNUM	0.9727	0.957	0.9648	0.9735	0.9212	0.9466	0.0001*
INITIALS	0.8333	0.3125	0.4545	0.7500	0.3750	0.5000	0.5024
INSTITUTION	0.9325	0.8526	0.8908	0.9330	0.8455	0.8871	0.2483
IPADDRESS	1	0.9	0.9474	1	0.9	0.9474	1
LOCATION	0.8738	0.6995	0.777	0.8616	0.6919	0.7675	0.6484
NAME	0.9447	0.9456	0.9452	0.9326	0.9100	0.9212	0.0001*
OTHER	0.8468	0.7387	0.7891	0.8528	0.7323	0.7880	0.6884
PHONE	0.9442	0.9087	0.9261	0.9564	0.8756	0.9142	0.0774
SSN	1	0	0	1	0	0	1
All	0.9508	0.9192	0.9348	0.9494	0.9029	0.9256	0.0001*

P = precision; *R* = recall; *F* = *F*-measure.
* Indicates statistically significant values (*p*-value < 0.0038 with Bonferroni correction).

i2b2 corpus is very close to, but never higher than, when using the i2b2 corpus alone.

3.2.2.2. Performance on the PhysioNet corpus. When testing on the PhysioNet corpus but training only on the new CCHMC or the i2b2 corpora, performance is very low (overall *F*-measures of 0.3379 and 0.2786) compared to training on the PhysioNet corpus. However, compared with the i2b2 experiment, the new CCHMC experiment still achieved better *F*-measure that amounts to 21% improvement (Table 6). *F*-measure is slightly higher when combining the new CCHMC corpus and the PhysioNet corpus than when using the PhysioNet corpus alone (Fig. 3). Recall is always substantially higher (at least +0.03), while precision is lower. Using the new CCHMC corpus together with the PhysioNet corpus yields a better balance between recall and precision. Combining the i2b2 corpus with the PhysioNet corpus improves performance when using 10–30% of the PhysioNet corpus. After that threshold, results become lower or indistinguishable from when using the PhysioNet corpus alone.

3.2.2.3. Performance on the original CCHMC corpus. When testing the de-identification performance on the original CCHMC corpus, performance is low when training only with the i2b2 corpus or the PhysioNet corpus (overall *F*-measures of 0.4689 and 0.2237). Performance is excellent when the model is trained on the original

Table 6
Performance on the i2b2 (a), PhysioNet (b), and original CCHMC (c) corpora when training on one corpus and testing on another (italic lines indicate same-corpus training and testing for comparison purposes).

Trained on	Nb of training instances	*P*	*R*	*F*
(a) Performance on i2b2				
new CCHMC	27,723	0.6367	0.3731	0.4705
PhysioNet	1771	0.2729	0.1091	0.1558
i2b2	*14,253*	*0.9682*	*0.9342*	*0.9509*
(b) Performance on PhysioNet				
new CCHMC	25,245	0.5570	0.2425	0.3379
PhysioNet	*1596*	*0.9534*	*0.599*	*0.7358*
i2b2	17,098	0.4286	0.2064	0.2786
(c) Performance on original CCHMC				
Original CCHMC	*27,638*	*0.9508*	*0.9192*	*0.9348*
i2b2	14,253	0.6362	0.3712	0.4689
PhysioNet	1773	0.3457	0.1653	0.2237

CCHMC corpus (*F*-measure of 0.9348). Fig. 4 displays the performance when the i2b2 and PhysioNet corpora are combined with increasing percentages of the original CCHMC corpus. It shows that combining the i2b2 corpus with the original CCHMC corpus yielded equivalent and sometimes lower performance than when using the original CCHMC corpus alone. The same phenomenon is observed when combining the PhysioNet and original CCHMC corpora.

4. Discussion

Our first set of experiments showed that using the new CCHMC corpus with IRB requested modifications to train the machine learning system to de-identify original CCHMC data yielded high performance (*F*-measure of 0.9256). Although numerically somewhat lower, these results were still quite close to those obtained when the original corpus was used to train the same machine learning system (*F*-measure of 0.9348). Although the difference was statistically significant for IDNUM, NAME and overall the numerical differences in *F*-measure were very small. The largest difference was less than 2.5% *F*-measure decrease for NAME (0.024). This contrasts sharply with findings from Yeniterzi et al. whose training on a resynthetized corpus and testing on real data resulted in ten times larger drop in performance (0.728 vs. 0.960 *F*-measures) [11]. Our better results can be explained partly by the fact that our new corpus retains most of the original PHI structure (although shuffled through the corpus), as opposed to the corpus of Yeniterzi et al., which is fully resynthetized. Another reason is that while our corpus was manually annotated for PHI, the corpus of Yeniterzi et al. was automatically de-identified by software. Consequently Yeniterzi's resynthetized corpus might have contained errors that impacted the subsequent de-identification performance. We believe that the most likely explanation for our better results is the careful engineering process behind the replacement and shuffling of PHI.

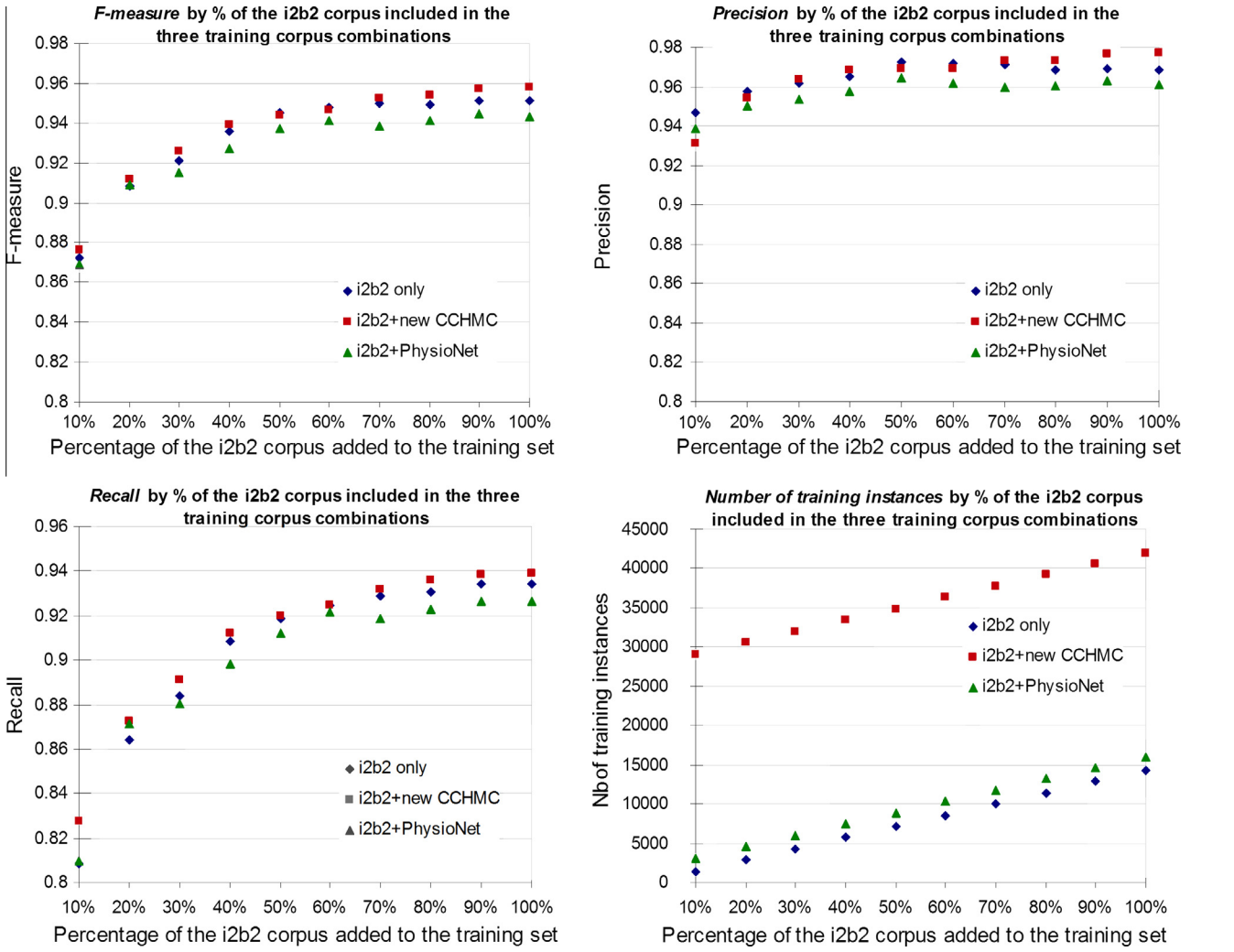

Fig. 2. De-identification performance (*F*-measure, precision, recall) on the i2b2 test corpus and number of training instances for models trained on various combinations of i2b2, new CCHMC and PhysioNet corpora, by percentage (horizontal axis) of the i2b2 corpus included in the training corpus.

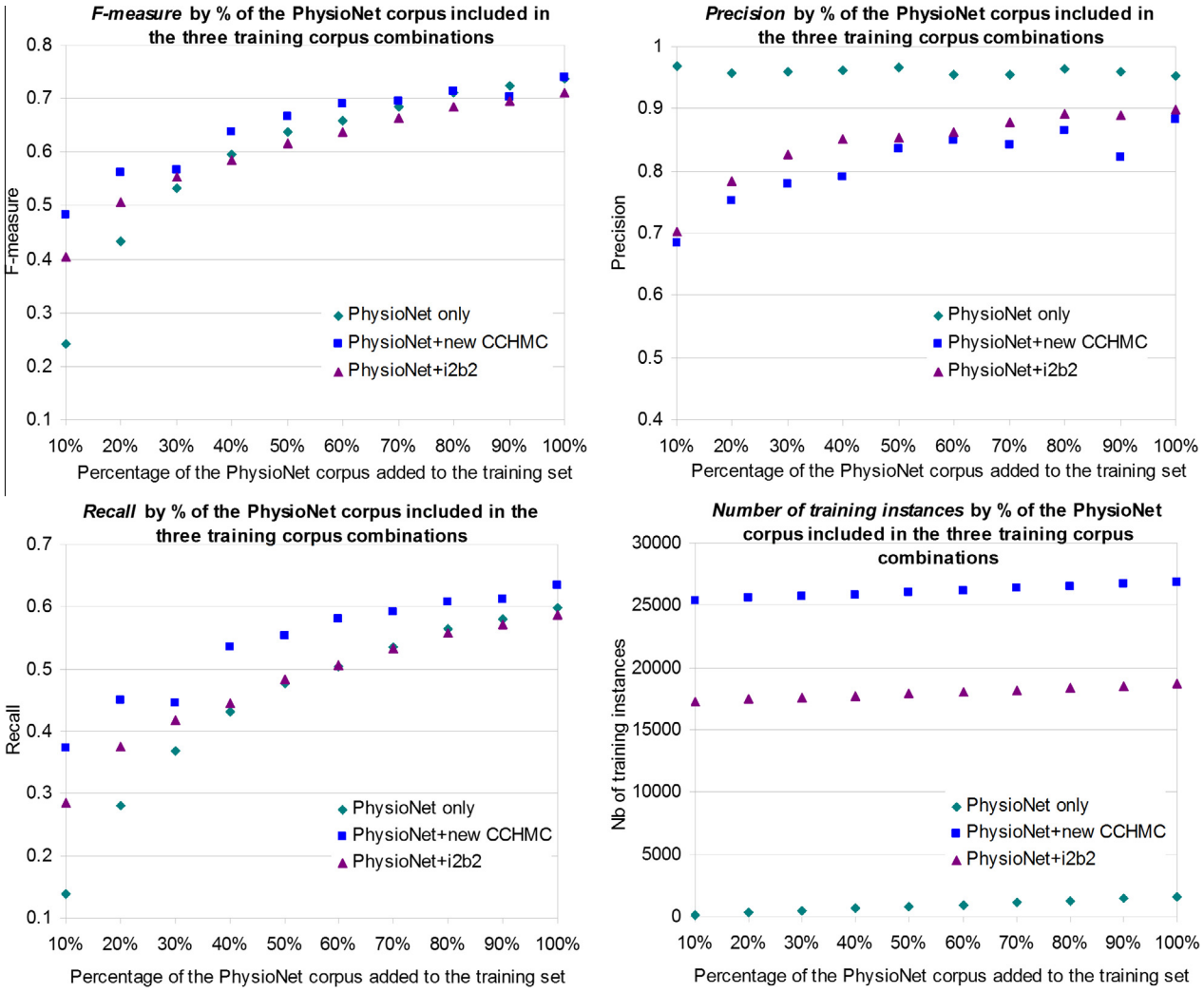

Fig. 3. De-identification performance (*F*-measure, precision, recall) on the PhysioNet corpus and number of training instances for models trained on various combinations of PhysioNet, new CCHMC and i2b2 corpora, by percentage (horizontal axis) of the PhysioNet corpus included in the training corpus.

Both Yeniterzi et al.'s replacement strategy and our method present advantages and drawbacks. Our approach retains more of the original corpus' research value. Yeniterzi et al.'s approach seems to present less risk of re-identification since it replaces all PHI by synthetic ones. However the risk is not completely eliminated. Because their automated system identifies PHI elements before replacing them, there is a higher risk of missed PHI elements remaining in the corpus. Overall, the risk of re-identification is very hard to quantify, since many factors can play a role (how rare a patient's condition is, how many personal (but non-protected) details are included in the notes such as family situation and employment).

The machine learning training and testing experiments on the two existing de-identification corpora and the CCHMC corpus showed that the new modified CCHMC corpus will contribute to training accuracy of de-identification tasks beyond the walls of our institution. Training on the new CCHMC corpus and testing on either the i2b2 or PhysioNet corpora resulted in much higher *F*-measures than training on either PhysioNet and testing on i2b2 (0.47 vs. 0.16) or training on i2b2 and testing on PhysioNet (0.34 vs. 0.28) corpora. Training on the new CCHMC corpus contributes more to the testing performance on either i2b2 or PhysioNet corpora than training on either the i2b2 or the PhysioNet corpora alone and testing on the original CCHMC corpus, although the training contribution of the i2b2 corpus is almost equal to the new CCHMC corpus' in cross training (Table 6). However, the experiments also showed that each corpus has its own unique characteristics that cannot be learned from the other two corpora. Indeed, when training on one corpus and testing on a different one, performance is always much lower than when training and testing on the same corpus. At this point the best cross training – training on the CCHMC corpus and testing on the i2b2 corpus – yielded a 0.47 *F*-measure, which is only 50% as effective as i2b2 in-corpus training (0.95 *F*-measure, Table 6). Hence, models trained on our corpus alone will only be partially transferrable to other institutions (depending on how different the other institution's data is from ours).

Nevertheless, we also showed in our corpus combination experiments that merging our corpus with a corpus from a different institution brought slightly better performance in de-identifying a test corpus from that institution than when only using this particular corpus on its own. Therefore, our corpus could reduce the amount of work and associated cost of building a high performance de-identification system in another institution. Using our corpus in combination with another corpus means that the other corpus might not need to be as large as it would if used on its own. Also, our corpus can be used to train a system and pre-annotate a new corpus to speed up the manual annotation process [24]. An interesting direction for future work is to investigate domain adaptation

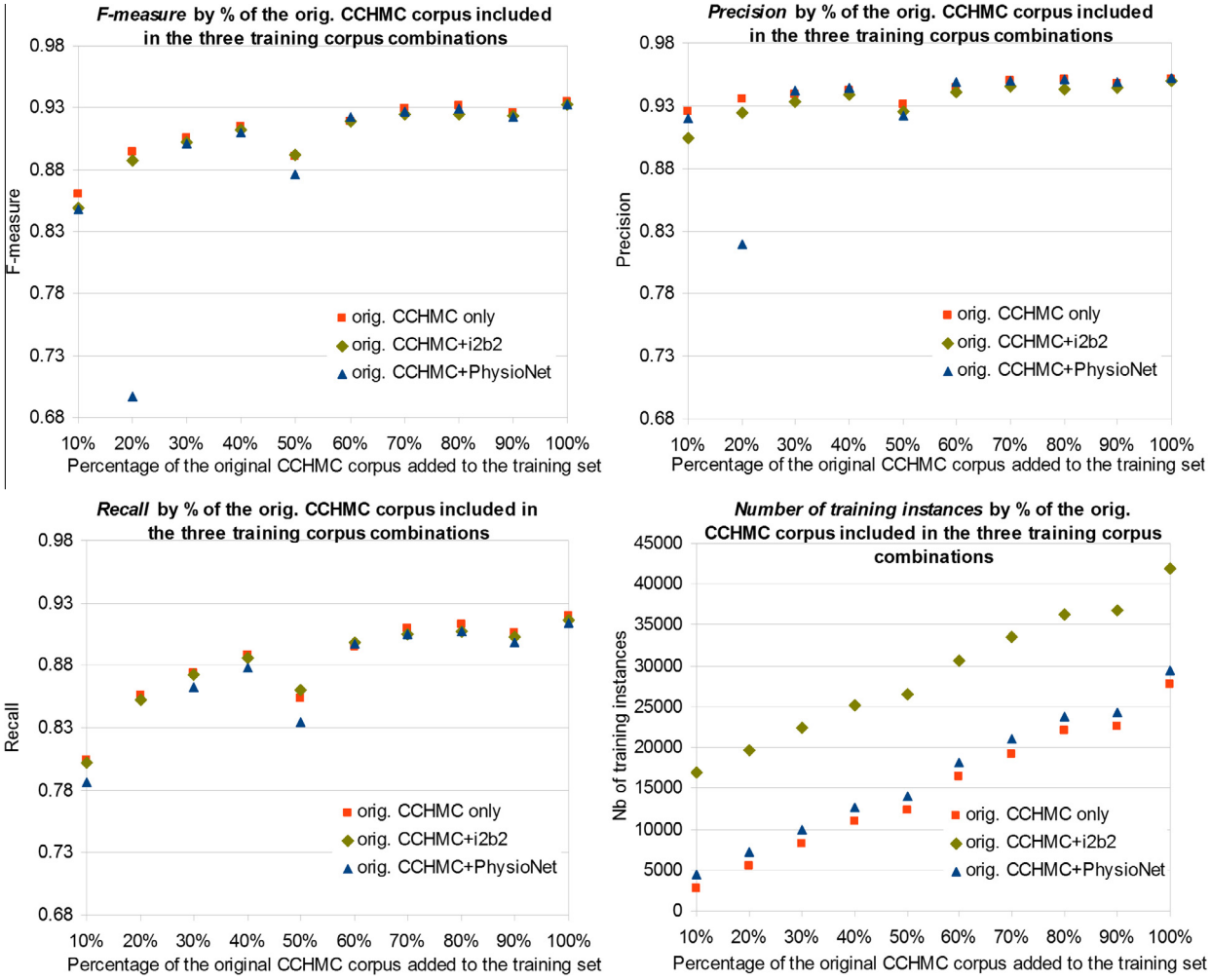

Fig. 4. De-identification performance (*F*-measure, precision, recall) on the original CCHMC corpus (orig. CCHMC = original CCHMC) and number of training instances for models trained on various combinations of original CCHMC, i2b2 and PhysioNet corpora, by percentage (horizontal axis) of the original CCHMC corpus included in the training corpus.

methods to improve the machine learning algorithms so that they are less dependent on the underlying dataset to achieve cross-corpora performance closer to in-corpus training experiments.

One of the limitations of the study is that the three corpora did not utilize the exact same annotation schema. Consequently, direct PHI level comparisons are not possible. Hopefully additional de-identification datasets will be released in the future and this will allow cross-training comparisons on the PHI level.

We can also perform a mapping between the three annotation schema from the three corpora (CCHMC, i2b2, and PhysioNet) to obtain a larger joint multi-institution corpus.

5. Conclusion

We successfully modified an existing in-house developed annotated gold standard de-identification dataset for external sharing per the requests of the IRB while preserved the de-identification research value of the modified corpus. The carefully engineered modifications caused only a limited drop in machine learning de-identification performance unlike an earlier published effort of a different group. The cross-training experiments verified that training on the new corpus contributes more to the testing performance than either training on two existing de-identification corpora alone, implying the valuable research value of the new corpus to the biomedical research community.

Based on the lessons learned during this work we suggest that researchers facing the task of boosting a clinical corpus by a more abundant corpus should:

1. harmonize the annotation sets: group or extend the annotation classes if needed;
2. use the large corpus to pre-annotate the smaller corpus before manual annotation;
3. annotate the smaller corpus in several iterations, each time re-training and re-evaluating performance;
4. evaluate when training with and without using the larger corpus.

Acknowledgments

The authors and annotators were supported by internal funds from Cincinnati Children's Hospital Medical Center. IS, LD and TL were partially supported by Grants 5R00LM010227-05, 1R21HD072883-01, and 1U01HG006828-01.

We thank the anonymous reviewers whose comments helped improve and clarify the manuscript.

References

[1] 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].
[2] Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32 [Epub 2008/07/26].
[3] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc: JAMIA 2007;14(5):550–63 [Epub 2007/06/30].
[4] Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak 2006;6:12 [Epub 2006/03/07].
[5] Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol 2004;121(2):176–86 [Epub 2004/02/27].
[6] Thomas SM, Mamlin B, Schadow G, McDonald C. A successful technique for removing names in pathology reports using an augmented search and replace method. In: Proceedings of the AMIA annual symposium; 2002. p. 777–81 [Epub 2002/12/05].
[7] Morrison FP, Li L, Lai AM, Hripcsak G. Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? J Am Med Inform Assoc: JAMIA 2009;16(1):37–9 [Epub 2008/10/28].
[8] Benton A, Hill S, Ungar L, Chung A, Leonard C, Freeman C, et al. A system for de-identifying medical message board text. BMC Bioinformatics 2011;12(Suppl. 3):S2 [Epub 2011/06/17].
[9] Gardner J, Xiong L. HIDE: an integrated system for health information DE-identification. Comp Med Syst 2008:254–9.
[10] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, et al. PhysioBank, PhysioToolkit, and Physionet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 2000;101(23):e215–20.
[11] Yeniterzi R, Aberdeen J, Bayer S, Wellner B, Clark C, Hirschman L, et al. Effects of personal identifier resynthesis on clinical text De-identification. J Am Med Inform Assoc 2010;17(2):159–68.
[12] Alfalahi A, Brissman S, Dalianis H. Pseudonymisation of personal names and other PHIs in an annotated clinical Swedish corpus. In: Third workshop on building and evaluating resources for biomedical text mining workshop programme; 2012.
[13] 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):849–59 [Epub 2010/10/19].
[14] Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2012. doi:10.1136/amiajnl-2012-001012 [August 2, Epub ahead of print].
[15] Deleger L, Li Q, Lingren T, Kaiser M, Molnar K, Stoutenborough L, et al. Building gold standard corpora for medical natural language processing tasks. In: American medical informatics association annual symposium proceedings; 2012. p 144–53.
[16] US Census Bureau. http://www.census.gov/genealogy/names/.
[17] Sibanda T, Uzuner O. Role of local context in automatic deidentification of ungrammatical, fragmented text. In: Proceedings of the main conference on human language technology conference of the North American chapter of the association of computational linguistics; 2006. p. 65–73.
[18] McCallum AC. MALLET: a machine learning for language toolkit; 2002.
[19] Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. Methods Inf Med 1998;37(4–5):334–44 [Epub 1998/12/29].
[20] Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc: JAMIA 2005;12(3):296–8. doi: 10.1197/jamia.M1733 [Epub 2005/02/03, PubMed PMID: 15684123; PubMed Central PMCID: PMC1090460]..
[21] Noreen EW. Computer-intensive methods for testing hypotheses: an introduction. New-York: Wiley; 1989.
[22] Cinchor N. The statistical significance of MUC4 results. In: MUC4 '92 Proceedings of the 4th conference on message understanding; 1992. p. 30–50.
[23] Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310(6973):170 [Epub 1995/01/21. PubMed PMID: 7833759; PubMed Central PMCID: PMC2548561]..
[24] Lingren T, Deléger L, Molnar K, Zhai H, Meinzen-Derr J, Kaiser M, et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: NLP gold standard development for clinical named entity recognition in clinical trial announcements. J Am Med Inform Assoc 2013. http://dx.doi.org/10.1136/amiajnl-2013-001837 [September 3, Epub ahead of print].