Published in final edited form as:

Int J Med Inform. 2014 October ; 83(10): 750–767. doi:10.1016/j.ijmedinf.2014.07.002.

De-identification of clinical narratives through writing complexity measures

Muqun Li^{a,*}, David Carrell^b, John Aberdeen^c, Lynette Hirschman^c, and Bradley A. Malin^{a,d}

^aDepartment of Electrical Engineering & Computer Science, Vanderbilt University, Nashville, TN, United States

^bGroup Health Research Institute, Seattle, WA, United States

^cThe MITRE Corporation, Bedford, MA, United States

^dDepartment of Biomedical Informatics, Vanderbilt University, Nashville, TN, United States

Abstract

Purpose—Electronic health records contain a substantial quantity of clinical narrative, which is increasingly reused for research purposes. To share data on a large scale and respect privacy, it is critical to remove patient identifiers. De-identification tools based on machine learning have been proposed; however, model training is usually based on either a random group of documents or a pre-existing document type designation (e.g., discharge summary). This work investigates if inherent features, such as the writing complexity, can identify document subsets to enhance de-identification performance.

Methods—We applied an unsupervised clustering method to group two corpora based on writing complexity measures: a collection of over 4500 documents of varying document types (e.g., discharge summaries, history and physical reports, and radiology reports) from Vanderbilt University Medical Center (VUMC) and the publicly available i2b2 corpus of 889 discharge summaries. We compare the performance (via recall, precision, and *F*-measure) of de-identification models trained on such clusters with models trained on documents grouped randomly or VUMC document type.

Results—For the Vanderbilt dataset, it was observed that training and testing de-identification models on the same stylometric cluster (with the average *F*-measure of 0.917) tended to outperform models based on clusters of random documents (with an average *F*-measure of 0.881). It was further observed that increasing the size of a training subset sampled from a specific cluster could yield improved results (e.g., for subsets from a certain stylometric cluster, the *F*-measure raised from 0.743 to 0.841 when training size increased from 10 to 50 documents, and the *F*-

© 2014 Published by Elsevier Ireland Ltd.

*Corresponding author at: Department of Biomedical Informatics, 2525 West End Avenue, Suite 1030, Nashville, TN 37205, United States. Tel.: +1 615 975 1328. muqun.li@vanderbilt.edu (M. Li).

Author contributions: M.L. designed the framework, performed the study, analyzed the experiment results and wrote the paper. D.C. assisted in the experiment design, data analysis, and revised the paper. J.A. and L.H. provided insights in the outline of the study, analysis of the results and the refinement of the paper. B.M. designed and supervised the whole study, analyzed the data, and revised the paper.

Competing interest: No conflict of interest exists in this paper.

measure reached 0.901 when the size of the training subset reached 200 documents). For the i2b2 dataset, training and testing on the same clusters based on complexity measures (average *F*-score 0.966) did not significantly surpass randomly selected clusters (average *F*-score 0.965).

Conclusions—Our findings illustrate that, in environments consisting of a variety of clinical documentation, de-identification models trained on writing complexity measures are better than models trained on random groups and, in many instances, document types.

Keywords

Electronic medical records; Privacy; Natural language processing

1. Introduction

Electronic health record (EHR) systems have been adopted to capture the condition of patients, facilitate communication between healthcare providers [1,2], and improve the quality of care [3]. While certain aspects of patient care are documented by structured data (e.g., diagnosis codes [4]) or semi-structured [5] text (e.g., problem lists [6]), a significant portion of documentation is captured only in the clinical narrative [7]. The narrative in the medical record has proven to be a great enabler of flexibility in clinical workflow [8] and decision support [9]. It has also been instrumental in supporting the extraction of complex phenotypic phenomena that lack explicit representation in coded data [10–12]. Nonetheless, it remains a challenge for healthcare organizations to share free text for research purposes, and concerns over the privacy of patients to which the records correspond is one of the primary reasons [13].

The need to protect patient privacy is supported by regulation in various countries. In the United States, for instance, the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) [14] limits the utilization of sensitive patient information (sans consent) without obscuring the identity of the patient and their relatives. There are various ways to meet the de-identification requirements of the Privacy Rule, but the most common is the Safe Harbor standard, which states that health information must be devoid of eighteen types of protected health information (PHI). These include explicit identifiers (e.g., patient names) and quasi-identifiers (e.g., dates).

Manual de-identification of clinical narratives can be costly, both in labor and time [15], making it near impossible to construct a corpus of more than a few hundred documents. This is clearly not scalable because the EHRs of even moderate size health care institutions can accumulate millions of electronic documents per year. For this reason, automated de-identification tools have been developed to make the process more efficient and replicable. These tools turn the de-identification task into a natural language processing (NLP) problem and can be roughly categorized into three groups: (1) rules and dictionaries, (2) machine learning, and (3) a hybrid of the two [16]. It has been demonstrated that rule-based systems perform well when informed by local knowledge and hand-crafted rules. But the time necessary to elicit such knowledge is a significant barrier to scalability and portability. By contrast, machine learning solutions tend to be more generalizable and robust because they *infer* de-identification models from numerous textual features derived automatically from

annotated training data. Nevertheless, machine learning approaches require a certain amount of manually annotated narrative to inform the learning process. Hybrid models, which strive to integrate the best of both rule-based and machine learning-based algorithms, can improve de-identification performance, but require both local knowledge and annotated training data [16].

This paper is primarily concerned with the scalability challenge in de-identification systems based on machine learning. It has been shown that training and testing on clinical narratives of the same document type (e.g., models trained on, and subsequently applied to, discharge summaries) yield the best performance [17]. Yet, such document type information is not always available and may not, in fact, provide the best basis, for grouping records into training classes due to heterogeneity in documentation practices. As such, it has been shown that training on a random selection of documents across the enterprise may allow for adequate performance. However, we hypothesize that mathematically calculable characteristics of the clinical documents themselves, specifically writing complexity and richness of clinical vocabulary, can be used to improve the performance of machine learned de-identification models by more effectively classifying documents into more homogeneous groups for model training and de-identification. Such methods may be of particular value for de-identifying corpora containing highly diverse narratives. To assess this hypothesis, we developed a feature extraction and clustering strategy to partition clinical documents into inferred types (categorized by writing complexity and clinical vocabulary usage) over which de-identification models are trained and tested.

We evaluate this hypothesis using two corpora. The first consists of the 889 discharge summaries from the i2b2 challenge. The second corpus consists of over 4500 medical records from the Vanderbilt University Medical Center (VUMC). Specifically, we investigate three alternative scenarios: document clustering by (1) EHR-assigned document type, (2) writing complexity and clinical vocabulary richness, and (3) a random process.

2. Background

2.1. Machine learning and de-identification tools

There are various machine learning approaches to de-identification that have been developed. These include strategies based on maximum entropy models [18], decision trees [19], support vector machines [20], and conditional random fields [21,22] (CRF). CRFs [23], in particular, have been broadly applied by the NLP community to solve various problems, such as shallow parsing in sequence labeling tasks [24] and biomedical named entity recognition [25]. In the context of de-identification, the task is generalized to a named entity tagging problem [17], such that the goal is to identify and correctly assign type labels to each PHI instance (e.g., person names, ages, and calendar dates). As a brand of classifier designed to label words according to such types, CRFs presume that dependencies exist between these type labels, and then capture these dependencies under a first-order Markov assumption.

Various software tools have adopted CRFs for de-identification. The Health Information DE-identification (HIDE) [22] and a tool at Cincinnati Children's Hospital [2] were both

developed based on the Mallet toolkit [26], while the Best-of-Breed (BoB) system [27] incorporates a CRF implementation from the Stanford NLP group [27,28]. For this study, we work with the MITRE Identification Scrubber Toolkit (MIST), which is based on the Carafe toolkit [29]. We use MIST because its built-in functionality addresses several useful tasks, including a web-based graphical annotation interface, a tagging module, a redaction and re-synthesis module, and an automated experiment engine.

2.2. Writing styles

Writing styles are author-specific literary patterns [30], the formal characterization of which can be traced back to at least the 19th century [31]. A wide array of assessment functions has been proposed to quantitatively represent personal writing style in the form of stylometric features. These measurements vary in their nature; e.g., lexical features are based on word-tokens [31], while another feature class is derived from syntactic information, approximated by part-of-speech [32].

In this work, we utilize two general types of writing complexity. The first type corresponds to *readability*. A readability formula is a measure designed to provide quantitative estimates of the difficulty in writing style [33,34]. Such formulae tend to count language variables in a document and estimate the reading difficulty level based on the associated statistics [34]. The earliest formulation of this concept [35] is attributed to Lively and Pressey for children's readability,³ which measured the vocabulary difficulty with specific indicators. This work influenced later readability studies, notably the Flesch's Reading Ease formula [36] (one of the most widely used readability measurements in the world [33]), which predicts the ease of reading based on ratios of syllables per word and words per sentence. Over 200 readability formulas have been published [35], and in this work we focus on several representative formulae, which are presented in Table 2.

Our second type of writing complexity is *lexical richness* (or *diversity*). This is a subclass of lexical features based on vocabulary size, frequency distribution, and other variations [30]. Lexical richness include a simple type token ratio (TTR):

$$\text{TTR}(N) = \frac{V(N)}{N},$$

where N denotes a document's word count and $V(N)$ denotes the number of unique words.

Some measures are variations of TTR, such as Carroll's Corrected TTR [37]. Others, such as Dugast's Uber Index [38], incorporate simple transformations of $V(N)$ and N .

3. Methods

3.1. Materials

In this study, we applied our methodology to two datasets.

For the first dataset, we selected patient records from an existing de-identified version of StarChart, the VUMC EHR [39]. The system contains information dating back to 1984 and continues to receive feeds from a diverse set of sources, including lab results, radiology reports, and external transcription companies. In lieu of a large human annotated corpus, we leveraged a locally specialized version of DE-ID, which is a commercially available rules-based de-identification software tool [40], to indicate the position and syntactic type (e.g., name vs date) of each patient identifier. DE-ID replaces identifiers with generic placeholders for the syntactic type. This decision was based on the fact that currently DE-ID, in conjunction with several pre- and post-processing modules developed at the VUMC, remains one of the core technologies by which the VUMC de-identifies its medical records (over 1.7 million) for local investigator-initiated research projects [41]. In addition, in a demonstration project conducted at the VUMC for its institutional review board, it was shown that the software in place at the VUMC exhibited a recall of over 99.9% for HIPAA Safe Harbor identifiers.

The corpus used for this study was randomly selected across dates and portions of the clinical enterprise. It consists of 4597 clinical documents of 9 EHR types, including discharge summaries (DS), history & physical assessments (HP), radiology notes (RAD), and pathology notes (PATH). The frequency distribution of documents by EHR type is shown in Table 1. There are approximately 550 documents of each type, except for family histories (FH) which consisted of 200.

The second dataset corresponds to the i2b2 de-identification challenge corpus [17]. It contains 889 annotated clinical notes in the form of hospital discharge summaries from the Partners Healthcare System, within which the real identifiers were replaced by synthetic information.

3.2. Document pre-processing and clustering based on complexity and richness measures

The process by which documents were pre-processed and clustered is depicted in Fig. 1, which also provides an overview of the experimental design. In general, the process of stylometric clustering corresponds to three stages: (1) Data Preprocessing (resynthesis of DE-IDed documents), (2) Feature Extraction, and (3) Unsupervised Document Clustering.

3.3. Preprocessing

The DE-ID software tool in use at VUMC replaces instances of PHI with fillers of the corresponding type (e.g., “John Smith could be replaced by “**NAME[XXX WWW]”). To correctly calculate the writing style features and compare results with the i2B2 corpus, we transformed DE-ID'ed documents into realistic text because the fillers might interfere with the syntax or semantics of the sentences, and thus impact the complexity score computations. Realistic text was also needed for training the machine-learned models that would subsequently be used to evaluate the impact of alternative clustering strategies on de-identification performance. The conversion from de-identified documents was accomplished using the MIST Resynthesis Engine [17]. This software engine creates plausible replacements for the PHI fillers, generating documents looking realistic to users [42]. A fictional example of a clinical document pre- and post-resynthesis is depicted in Fig. 2.

3.3.1. Complexity and richness feature extraction—The goal of complexity and richness feature extraction is to construct an appropriate feature set for each clinical document. This set is then translated into a corresponding feature matrix supplied to a machine learning algorithm, where each row and column of the matrix corresponds to an individual document and complexity feature value, respectively. For this work we used three types of measures to characterize document writing style. The first two correspond to the readability and lexical richness measures mentioned earlier. The implementation for the reading complexity measures and lexical richness measures were accomplished using the Korpus package in the R programming environment [43].

The third writing style measure we refer to as *clinical term richness*. Here, we justify this measure and describe its computation. While the two types of writing complexity we use have been in use for decades, they only characterize the complexity of a document's syntax. Our documents all derive from the clinical domain, but a wide range of healthcare workers compose them (e.g., nurses, technicians, physicians from various medical specialties). We hypothesized that the characteristics of PHI in documents may correlate with the quantity of clinical language they contain, and that clinical language “richness” would be an important factor to consider when clustering documents for de-identification purposes. The clinical term richness measure was thus designed to represent the ratio of the sum of all clinical terms in a document to the sum of all of its terms, as shown in Table 2. To identify clinical terms we processed each document with MetaMap [44], which maps words and phrases to the Unified Medical Language System's (UMLS) Metathesaurus [45]. This clinical richness ratio was included with the writing complexity measures in each document's feature vector.

We started with 19 readability formulas and 12 lexical richness measures. To eliminate the effect caused by correlations between measures, we performed a feature selection process based on Regularized Random Forest (RRF) on the 31 features [49]. The process was performed independently for each of the corpora. The final feature sets for the VUMC and i2b2 datasets are shown in Table 2. The feature set for the VUMC corpus was composed of 8 measures: 5 associated with readability measurement, 2 with lexical richness assessment, and the clinical term richness measure. The fact that the latter measure was retained indicated that this feature was sufficiently distinct from standard writing complexity measures. The feature set for the i2b2 dataset was composed of 9 measures, again including the clinical term richness measure, among which 6 were related to readability measures, 2 with lexical richness.

3.3.2. Unsupervised document clustering—Clustering was performed in an unsupervised manner according to a hierarchical model, with Ward's coefficient [50] as the measure of goodness of fit. The number of clusters was determined by evaluating the quality of the clusters based on the Dunn index [51]. Using this criterion, we assigned each of the documents in our study corpus of the VUMC dataset to one of 13 clusters, which ranged in size from 60 to 672 documents. The i2b2 data was partitioned into 2 clusters, the sizes of which were 413 and 476, respectively.

4. Experimental design

4.1. Training and testing data preparation

To perform our experiments, we defined the following sets of clusters for the Vanderbilt data:

1. *VUMC EHR-assigned document types*. This set of clusters was based on document type assigned to each document as it was composed in the EHR by a VUMC employee (Table 1). We refer to these clusters as $T_1, \dots, T_9$.
2. *Complexity measures*. This set of clusters was based on the process described in the previous section. We refer to these clusters as $C_1, \dots, C_{13}$.
3. *Complexity measure-based subsets*. These sets were randomly selected from clusters $C_1, \dots, C_{13}$, with varying size over $\{10, 20, 30, 50, 75, 100, 150, 200\}$, denoted as $Cs_{i,j}$, from which i ranges from 1 to m and j corresponds to the number of randomly selected documents (e.g., $Cs_{i,10}$ corresponds to a set of 10 randomly selected documents from the i th cluster).
4. *Random processes*. We generated random groups of equal size for the stylometric clusters to ensure this type of cluster contains a comparable amount of data with the complexity-based clusters. In doing so, we derived a third set of clusters $R_1, \dots, R_{13}$, such that $|C_i| = |R_i|$ for all clusters.
5. *Large random clusters*. After generating clusters from complexity measures, for each cluster C_i , we gather all the remaining clusters (i.e., $C_1, \dots, C_{i-1}, C_{i+1}, \dots, C_{13}$) to create a large random document set. We derived several random subsets from such large clusters, with varying size over $\{100, 200, 500, 1000\}$. These random subsets are denoted as $D_{i,j}$, where i ranges from 1 to m and j corresponds to the number of randomly selected documents (e.g., $D_{i,100}$ corresponds to 100 randomly selected documents not in the i th cluster). The goal of experiments based on these subsets is to create a baseline for the evaluation of stylometric clusters.

For evaluation, we performed five sets of experiments using 10-fold cross-validation in each scenario. First, we trained and tested over each of the datasets, $\{C_1, \dots, C_{13}\}$, $\{R_1, \dots, R_{13}\}$, and $\{T_1, \dots, T_9\}$ (as shown in Figs. 1 and 3). Second, within the stylometric and random clusters, we evaluated the performance of the learned models in a cross-dataset manner (as shown in Fig. 1). For instance, we train a model over C_i and test it over $C_1, \dots, C_{i-1}, C_{i+1}, \dots, C_m$. Third, we trained on each stylometric cluster and tested with the corresponding same-size random set. Fourth, we evaluated the training performance of the large random clusters when testing on the stylometric clusters (as shown in Fig. A1 of online Appendix A). Specifically, we trained on each of the random subsets $D_{i,1}, \dots, D_{i,s}$ and tested against C_i . For the final set of experiments, which aim to evaluate the training sensitivity of the stylometric clusters, we trained and tested over all the subsets of such clusters, $\{Cs_{1,10}, \dots, Cs_{m,k}\}$ (as shown in Fig. A2 of online Appendix A).

4.2. De-identification model training

Separate de-identification models were trained for each cluster of documents using default MIST settings. The resulting models were used in their respective experiments to evaluate de-identification performance.

4.3. Evaluation measures

We report our results for each of the experiments using standard information retrieval measures. In the context of this study, these measures are defined as follows:

$$(P)_{\text{recision}} = \frac{\text{Number of PHI instances correctly labeled}}{\text{Number of PHI instances labeled by the system}}$$

$$(R)_{\text{ecall}} = \frac{\text{Number of PHI instances correctly labeled}}{\text{Number of PHI instances labeled in the gold standard}}$$

$$(F) - \text{measure} = \frac{2pr}{p+r}$$

5. Results

To provide insight into the composition of stylometric clusters, Fig. 4 reports on the distribution of document types for each cluster. These charts show that 9 out of 13 stylometric clusters are highly heterogeneous mixtures of VUMC document types, which confirms that our clustering method is not merely replicating the grouping of the document types. The clusters that were relatively homogenous in document type consisted of general reports (C_{12} and C_{13}), letters (C_{10}), and pathology notes (C_{11}).

Tables 3 and 4 summarize the evaluation results for the various de-identification models with the Vanderbilt dataset. For brevity, Tables 3 and 4 focus on results of the Vanderbilt data with respect to the F -measure. In general, recall and precision were balanced (the full results including recall and precision are presented in online Appendix B). We would like to note that our evaluation used the DE-ID processed documents as a gold standard, which indicates the impact of the DE-ID performance on our results. Table 5 presents the total number of PHI types and instances in training de-identification models for stylometric clusters.

5.1. De-identification training on complexity and random clusters

In Table 3, each row represents a test dataset. Since we tested models trained on each of the clusters, for each column, we report the best (*max*) and worst (*min*) F -measures with corresponding training datasets. We also report the F -measure of the model trained on the corresponding testing cluster (*self*). The results presented in the first half of the table were

generated by the scenario in which training and testing was performed with clusters derived from the complexity measures, while the second half correspond to random clusters.

Regarding the stylometric experiment, generally speaking, in seven of the clusters (C_1 , C_3 , C_4 , C_9 , C_{10} , C_{12} , C_{13}), training and testing on the same cluster (i.e., C_i vs C_i) performed better than testing with models based on different clusters (i.e., C_i vs C_j , $i \neq j$). For example, when testing on C_{13} , which is the smallest (60 documents), the model trained on documents from this cluster achieved an F -measure of 0.990, but the F -measure never reached higher than 0.808 for C_3 with models based on other clusters, in which cases the F -measure dropped as far as 0.101 for C_{12} . The possible explanation could be found in the limited number of PHI types and relatively sufficient number of PHI instances in the cluster, as shown in Table 5. It also suggests that training and testing on the same cluster might eliminate the limitation of insufficient training data. Specifically, training on smaller clusters (e.g., C_{12} , C_{13}) performed poorly when evaluating on clusters other than themselves, but exceptionally well when testing on their own.

The pattern found in the previous tests, however, did not hold true in the scenario where training and testing was performed with clusters derived from a random process. The results found in these experiments confirmed our hypothesis that training and testing on the same random clusters would not necessarily yield the best performance comparing with training and testing on different clusters based on random process. Rather, de-identification models trained on different clusters showed similar variations in testing, despite the fact that they were testing on their own or on other random clusters. Specifically, models created by R_{13} always performed worse than other training models (with the F -measure ranging from 0.775 to 0.869) including testing on itself. R_7 on the other hand was the best training set, whose F -measures dominated 9 out of all 13 testing sets, including testing on its own. Unlike the high variation in experiments for complexity-measure-based clusters, the F -measure was bounded between 0.775 and 0.959 in this half of Table 3.

5.2. De-identification training and testing crossing cluster types

Table 4 reports on the evaluation of experiments crossing cluster types. In the columns under $C \Rightarrow R$ we trained on stylometric clusters and tested on random clusters of the same size, while in $R \Rightarrow C$ we reversed the process. Comparing Tables 3 and 4, it can be seen that the stylometric models performed worse than random models on the random clusters, with an F -measure between 0.172 and 0.884. When testing on stylometric clusters, nine of the models trained on the testing clusters themselves performed better than random training models. R_2 , R_5 , R_7 and R_{11} achieved better models than their stylometric counterparts C_2 , C_5 , C_7 and C_{11} .

However, recalling the fact that six of the stylometric clusters showed best testing results with training models of clusters other than themselves, the four clusters (C_2 , C_5 , C_7 and C_{11}) were all included. When testing on C_7 and C_{11} , models trained on C_4 yielded the best performance over all stylometric clusters, with F -measures of 0.941 and 0.993, respectively, which also surpassed the best performances of models by random clusters R_7 and R_{11} . For testing on C_2 and C_5 , models trained on C_3 and C_6 yielded the best F -measures (0.881 and

0.925, respectively), both on par with the random models R_2 and R_5 (F -measures of 0.893 and 0.929, respectively).

5.3. De-identification training on large random clusters

The average performance for de-identification models trained on different sized subsets of large random-process based clusters and tested on complexity-measure based clusters is depicted in Fig. 5 (left). Figures depicting the performance of each of the large random clusters are included in Appendix C. Overall, larger size training sets performed better than smaller size sets, as shown in the left of Fig. 5. For instance, the F -measure of 100 documents for $D_{10,100}$ was 0.733, which increased with the sample size, such that by 1000 documents, the F -measure was 0.804.

In comparison to the results from Table 3, the F -measures for training and testing on a particular stylometric cluster C_i generally performed better than a sample size of 200 documents from D_i . It can be seen that 8 out of the 13 complexity clusters outperformed subsets of size 500 and five (C_4 , C_9 , C_{10} , C_{12} , and C_{13}) yielded better F -measures than the random subsets of 1000 documents. Specifically, C_{10} , C_{12} and C_{13} contain the least number of documents of all clusters – each with less than 200 documents. These findings indicate that for smaller clusters, training and testing on their own may achieve better performance than training models over a much larger dataset of randomly selected documents (while C_{11} was an exception to this trend).

5.4. De-identification training on subsets of complexity based clusters

The experimental results for de-identification models trained and tested on varying sizes of subsets from the complexity-measure based clusters are depicted in online Appendix C, for evaluating the scalability of such clusters. The right of Fig. 5 plots the average performance of stylometric clusters. Similar to the large random datasets, increasing sample size seems to improve de-identification model training.

Recalling the results from large mixed random clusters, a subset of C_{10} with the size of 75 already outperformed the corresponding random dataset of 1000 documents in terms of F -measure – the former achieved 0.813 while the latter achieved 0.804. With C_{12} and C_{13} , the 50-document subset yielded better F -measures than the 1000-document random clusters, a subset of 10 documents could achieve an F -measure of 0.961, while a random dataset of 1000 documents only achieved 0.859. Such findings suggest that smaller clusters tend to be homogeneous, for which training on a limited number of documents of their own group would surpass training on much larger random clusters.

5.5. De-identification training on clusters based on VUMC designated type

Beyond the statistics for VUMC-designated document type, Table 1 reports the testing results of training de-identification models based on these types. The overall performance was worse than the complexity-measures based clusters scenario. Here, it can be seen that documents of type CC generated the highest scores in all the measurements, with an F -measure of 0.964, and type FH produced the lowest F -measure of 0.842. Comparing all scenarios, the average F -measures of training and testing on the same cluster are ordered as

follows: complexity measures clustering (0.917) > document type clustering (0.911) > random clustering (0.881). The *F*-scores of the complexity-based clusters were significantly higher than those with random clusters ($t = 2.79$, $p = 0.006$); meanwhile, their *F*-score difference between the complexity clustering and document type based clustering were not statistically significant.

5.6. De-identification training on i2b2 data

Table 6 summarizes the experiment results on the i2b2 dataset. The training and testing procedures were consistent with the Vanderbilt dataset, except that the type-based clustering was not possible because the i2b2 data only contains discharge summaries.

For comparison, we randomly divided the dataset into 2 clusters (IR_1 and IR_2) with the same sizes as the stylometric clusters (413 and 476 documents).

It can be seen that, for the stylometric clusters, training and testing on the same cluster always yielded better *F*-scores than training and testing on different clusters. IC_2 performed best with an *F*-measure of 0.972. However, the general performance of both stylometric clusters was on par, with the random clusters. Specific, the stylometric and random cluster *F*-scores were 0.966 and 0.965, respectively. As such, there was no statistical significant difference.

However, the results are not entirely unsurprising. The findings suggest that the i2b2 dataset is a highly homogenous corpus and lacks the variance in writing styles that is amenable to our stylometric clustering strategy. This is consistent with the fact that the dataset consists entirely of discharge summaries from Partners Healthcare [52].

6. Discussion

There are several notable aspects of the experimental analysis to highlight. First, the results confirm the hypothesis that training on a stylometric cluster yields better de-identification performance when testing on the same cluster. Second, it was further confirmed that stylometric de-identification models yield results that are better than models based on random collections of documents (which is akin to training de-identification models on mixtures of many document types as is common done). One possible explanation of this observation is the writing style features lead to models that account for grammar and phrasing, which are critical in natural language processing. As a result, the models learned over these features could yield better performance than the ones based on random features. Third, the performance of stylometric models is, in many instances, better than those derived from VUMC-designated document types. In combination, these findings suggest that higher fidelity de-identification models can be composed with less training data and institutional knowledge. This is critical because there are potentially hundreds to thousands of document types generated in healthcare settings. Though the proposed strategy does not always yield the best de-identification model, we believe it is more scalable because of its hierarchical clustering strategy, which minimizes the number of clusters and thus de-identification models we apply. We note that hierarchical grouping could also be performed on document

types to reduce the number of total documents types, but this will still yield random groups due to the inability to partition predefined types.

Despite the positive nature of the results, we recognize that our study is limited in its scope for several reasons. First, we used a VUMC-specialized version of DE-ID as a proxy for a human-annotated corpus. This technology is oriented toward PHI recall and, thus may over-redact, which could explain why our F -measures are a little lower than those observed in gold standard environments. Second, our data was mainly based on the VUMC, a single healthcare organization. While there is data available from other sites, such as the i2b2 corpus, they tend to be composed solely of a single type (e.g., discharge summaries) and are relatively small in size (e.g., ~ 1500). The clustering experiment on the i2b2 data showed limited improvement on de-identification model training, indicating the homogeneity of i2b2 data obstructed the writing stylometric clustering. Third, the VUMC EMR contains over 1000 document types because EMR users are allowed to create custom document types, but we chose to focus on only nine of the larger size types in this setting to report on interpretable results and to avoid an overly complex analysis.

Finally, while our findings lend credibility to the claim that features based on readability and lexical richness can lead to better de-identification models than random groups of documents, we acknowledge that such features may be correlated with other unknown factors. For instance, though unlikely, the MIST resynthesis engine may embed behind a certain pattern in the new PHI instances. This pattern could, in turn, influence the grammar and phrasing. Second, the DE-ID process might induce a pattern in the documents as well. For instance, DE-ID obscures the distribution of character length of PHI instances, which could, in turn, bias the resynthesis process. Third, our findings are mainly based on documents from a single institution; i.e., the VUMC. It is possible that data from a different institution might not exhibit as robust a clustering or distinction in performance.

7. Conclusions

The performance of machine-learning based de-identification is largely determined by its training process. It has been shown that training and testing within the same document type (e.g., discharge summaries) can lead to better result than with a mixture of document types. Our study demonstrates that grouping the training data to be more homogenous in its features can yield better performance in machine learned de-identification models. Specifically in the scenario studied in this work, when document typing is unavailable, similar, if not more effective de-identification models can be derived by partitioning clinical documents into clusters based on their writing complexity (in terms of readability and lexical richness). The results are sufficiently strong to merit further investigation and integration into machine learning-based de-identification tools.

Acknowledgments

This research was sponsored, in part, by grants R01LM011366 and R01LM009989 from the National Library of Medicine, U01HG006385 and U01HG006378 from the National Human Genome Research Institute, and CCF-0424422 (Team for Research in Ubiquitous Secure Technology) from the National Science Foundation. The authors wish to thank Steve Nyemba from Vanderbilt University and Samuel Bayer, and Ben Wellner of the MITRE Corp. for helpful discussions during the formulation of this project.

Appendix A

This appendix provides an illustration of the experimental design for the scalability tests reported in the main manuscript.

The scenario depicted in Fig. A1 corresponds to the experimental setting in which de-identification models are trained on large random mixture clusters and tested on stylometric clusters. In this setting, we randomly sampled from the large mixture of random clusters, with sizes ranging from 100 to 1000. The models were trained on such random subsets and tested on complexity-based clusters.

The scenario depicted in Fig. A2 corresponds to the experimental setting in which de-identification models are both trained and tested on stylometric clusters. In this case, de-identification models were trained and tested on subsets of stylometric clusters, sizing from 10 to 200.

Fig. A1. Experimental design of the scalability test for large random mixture clusters

Fig. A2. Experimental design of the scalability test for complexity based clusters

Appendix B

This section contains complete results (in terms of precision, recall and F -measure) of the de-identification model training and testing experiments.

Due to page width limitations, results for complexity based clusters and random clusters are split into 3 tables each. Tables B1-a to B1-c correspond to clusters created by complexity

measures. Tables B2-a to B2-c are showing the results for random process generated clusters.

Table B1a
De-identification performance (in terms of Precision, Recall, and F-measure) based on clusters derived from the complexity measures

Train ⇒ ↓Test	<i>C</i> ₁			<i>C</i> ₂			<i>C</i> ₃			<i>C</i> ₄		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
<i>C</i> ₁ (<i>n</i> = 547)	0.922	0.921	0.921	0.839	0.806	0.821	0.899	0.913	0.906	0.902	0.918	0.910
<i>C</i> ₂ (<i>n</i> = 229)	0.877	0.886	0.879	0.892	0.815	0.851	0.873	0.891	0.881	0.868	0.874	0.871
<i>C</i> ₃ (<i>n</i> = 569)	0.943	0.935	0.939	0.873	0.842	0.857	0.940	0.947	0.944	0.937	0.939	0.938
<i>C</i> ₄ (<i>n</i> = 403)	0.842	0.845	0.843	0.740	0.674	0.705	0.830	0.840	0.835	0.867	0.876	0.872
<i>C</i> ₅ (<i>n</i> = 531)	0.923	0.925	0.924	0.870	0.805	0.836	0.897	0.918	0.907	0.914	0.927	0.921
<i>C</i> ₆ (<i>n</i> = 362)	0.849	0.852	0.851	0.759	0.689	0.721	0.826	0.859	0.842	0.860	0.883	0.871
<i>C</i> ₇ (<i>n</i> = 323)	0.937	0.942	0.940	0.891	0.828	0.858	0.936	0.939	0.937	0.935	0.947	0.941
<i>C</i> ₈ (<i>n</i> = 247)	0.881	0.876	0.878	0.809	0.725	0.764	0.852	0.867	0.859	0.881	0.892	0.887
<i>C</i> ₉ (<i>n</i> = 611)	0.872	0.858	0.865	0.790	0.749	0.768	0.819	0.860	0.839	0.846	0.878	0.862
<i>C</i> ₁₀ (<i>n</i> = 439)	0.766	0.772	0.769	0.674	0.586	0.625	0.749	0.766	0.758	0.815	0.832	0.823
<i>C</i> ₁₁ (<i>n</i> = 92)	0.988	0.970	0.978	0.947	0.878	0.910	0.976	0.985	0.980	0.989	0.997	0.993
<i>C</i> ₁₂ (<i>n</i> = 184)	0.935	0.834	0.882	0.757	0.758	0.757	0.844	0.837	0.840	0.839	0.873	0.855
<i>C</i> ₁₃ (<i>n</i> = 60)	0.994	0.569	0.722	0.978	0.542	0.687	0.993	0.681	0.808	0.997	0.610	0.755

Table B1b
De-identification performance (in terms of Precision, Recall, and F-measure) based on clusters derived from the complexity measures

Train ⇒ ↓Test	<i>C</i> ₅			<i>C</i> ₆			<i>C</i> ₇			<i>C</i> ₈		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
<i>C</i> ₁ (<i>n</i> = 547)	0.905	0.887	0.896	0.906	0.895	0.901	0.892	0.879	0.885	0.892	0.869	0.881
<i>C</i> ₂ (<i>n</i> = 229)	0.862	0.814	0.837	0.878	0.851	0.864	0.812	0.792	0.801	0.842	0.806	0.822
<i>C</i> ₃ (<i>n</i> = 569)	0.931	0.905	0.918	0.934	0.907	0.920	0.930	0.920	0.925	0.925	0.884	0.904
<i>C</i> ₄ (<i>n</i> = 403)	0.849	0.818	0.833	0.859	0.856	0.857	0.817	0.802	0.809	0.847	0.833	0.840
<i>C</i> ₅ (<i>n</i> = 531)	0.928	0.902	0.915	0.934	0.917	0.925	0.886	0.886	0.886	0.920	0.895	0.908
<i>C</i> ₆ (<i>n</i> = 362)	0.865	0.826	0.845	0.879	0.877	0.878	0.816	0.809	0.812	0.868	0.869	0.868
<i>C</i> ₇ (<i>n</i> = 323)	0.934	0.912	0.922	0.934	0.913	0.923	0.929	0.914	0.922	0.926	0.892	0.908
<i>C</i> ₈ (<i>n</i> = 247)	0.885	0.849	0.866	0.902	0.893	0.897	0.838	0.819	0.828	0.901	0.891	0.896
<i>C</i> ₉ (<i>n</i> = 611)	0.880	0.846	0.863	0.901	0.879	0.890	0.798	0.827	0.812	0.879	0.847	0.863
<i>C</i> ₁₀ (<i>n</i> = 439)	0.785	0.753	0.768	0.826	0.832	0.829	0.732	0.722	0.726	0.801	0.804	0.803

Train $\Rightarrow$ $\Downarrow$ Test	C_5			C_6			C_7			C_8		
	P	R	F	P	R	F	P	R	F	P	R	F
$C_{11} (n = 92)$	0.986	0.957	0.970	0.980	0.980	0.980	0.983	0.972	0.977	0.947	0.962	0.955
$C_{12} (n = 184)$	0.975	0.897	0.934	0.957	0.794	0.867	0.894	0.781	0.833	0.948	0.748	0.836
$C_{13} (n = 60)$	0.983	0.197	0.326	0.949	0.286	0.438	0.984	0.310	0.469	0.963	0.088	0.160

Table B1c
De-identification performance (in terms of Precision, Recall, and F-measure) based on clusters derived from the complexity measures

Train $\Rightarrow$ $\Downarrow$ Test	C_9			C_{10}			C_{11}			C_{12}			C_{13}		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
$C_1 (n = 547)$	0.894	0.865	0.879	0.877	0.856	0.866	0.344	0.601	0.433	0.796	0.732	0.762	0.115	0.347	0.172
$C_2 (n = 229)$	0.855	0.803	0.827	0.852	0.818	0.834	0.337	0.513	0.398	0.752	0.669	0.706	0.103	0.340	0.157
$C_3 (n = 569)$	0.925	0.878	0.901	0.911	0.855	0.882	0.382	0.629	0.466	0.819	0.768	0.792	0.136	0.367	0.198
$C_4 (n = 403)$	0.834	0.818	0.826	0.853	0.853	0.853	0.393	0.579	0.457	0.761	0.705	0.731	0.127	0.293	0.176
$C_5 (n = 531)$	0.929	0.907	0.918	0.905	0.908	0.907	0.418	0.653	0.495	0.847	0.794	0.819	0.095	0.358	0.150
$C_6 (n = 362)$	0.875	0.849	0.862	0.872	0.895	0.884	0.293	0.546	0.376	0.770	0.712	0.739	0.086	0.267	0.130
$C_7 (n = 323)$	0.929	0.890	0.909	0.906	0.875	0.890	0.417	0.614	0.492	0.814	0.796	0.805	0.114	0.345	0.169
$C_8 (n = 247)$	0.892	0.860	0.875	0.883	0.897	0.890	0.280	0.495	0.352	0.769	0.701	0.731	0.068	0.254	0.106
$C_9 (n = 611)$	0.912	0.899	0.906	0.854	0.871	0.862	0.350	0.639	0.446	0.793	0.745	0.768	0.079	0.302	0.125
$C_{10} (n = 439)$	0.780	0.764	0.772	0.883	0.895	0.889	0.290	0.524	0.364	0.667	0.613	0.638	0.105	0.256	0.148
$C_{11} (n = 92)$	0.973	0.986	0.979	0.972	0.980	0.976	0.963	0.957	0.960	0.922	0.921	0.920	0.378	0.718	0.492
$C_{12} (n = 184)$	0.925	0.860	0.891	0.820	0.787	0.803	0.459	0.557	0.492	0.973	0.981	0.977	0.199	0.331	0.246
$C_{13} (n = 60)$	0.551	0.156	0.243	0.965	0.194	0.319	0.394	0.062	0.107	0.897	0.054	0.101	0.994	0.987	0.990

Table B2a
De-identification performance (in terms of Precision, Recall, and F-measure) based on random clusters

Train ⇒ ↓Test	<i>R</i> ₁			<i>R</i> ₂			<i>R</i> ₃			<i>R</i> ₄		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
<i>R</i> ₁ (<i>n</i> = 547)	0.893	0.888	0.891	0.886	0.877	0.881	0.884	0.882	0.883	0.892	0.885	0.888
<i>R</i> ₂ (<i>n</i> = 229)	0.900	0.894	0.897	0.891	0.881	0.886	0.891	0.885	0.888	0.900	0.894	0.897
<i>R</i> ₃ (<i>n</i> = 569)	0.902	0.903	0.903	0.903	0.896	0.899	0.896	0.896	0.896	0.914	0.907	0.910
<i>R</i> ₄ (<i>n</i> = 403)	0.888	0.884	0.886	0.885	0.878	0.881	0.888	0.889	0.888	0.892	0.892	0.892
<i>R</i> ₅ (<i>n</i> = 531)	0.895	0.888	0.892	0.881	0.886	0.883	0.887	0.887	0.887	0.901	0.893	0.897
<i>R</i> ₆ (<i>n</i> = 362)	0.889	0.870	0.879	0.887	0.864	0.875	0.888	0.874	0.881	0.898	0.875	0.886
<i>R</i> ₇ (<i>n</i> = 323)	0.889	0.884	0.886	0.889	0.884	0.886	0.893	0.892	0.892	0.903	0.896	0.900
<i>R</i> ₈ (<i>n</i> = 247)	0.899	0.900	0.899	0.897	0.889	0.893	0.892	0.891	0.891	0.903	0.897	0.900
<i>R</i> ₉ (<i>n</i> = 611)	0.892	0.885	0.888	0.883	0.880	0.881	0.889	0.884	0.887	0.897	0.888	0.892
<i>R</i> ₁₀ (<i>n</i> = 439)	0.886	0.897	0.891	0.901	0.900	0.901	0.903	0.908	0.906	0.909	0.913	0.911
<i>R</i> ₁₁ (<i>n</i> = 92)	0.891	0.889	0.890	0.903	0.896	0.900	0.891	0.892	0.891	0.900	0.896	0.898
<i>R</i> ₁₂ (<i>n</i> = 184)	0.896	0.885	0.890	0.895	0.883	0.888	0.886	0.870	0.878	0.901	0.887	0.893
<i>R</i> ₁₃ (<i>n</i> = 60)	0.944	0.947	0.945	0.941	0.954	0.947	0.946	0.958	0.952	0.948	0.959	0.953

Table B2b
De-identification performance (in terms of Precision, Recall, and F-measure) based on random clusters

Train ⇒ ↓Test	<i>R</i> ₅			<i>R</i> ₆			<i>R</i> ₇			<i>R</i> ₈		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
<i>R</i> ₁ (<i>n</i> = 547)	0.885	0.882	0.883	0.881	0.882	0.881	0.894	0.895	0.894	0.890	0.877	0.883
<i>R</i> ₂ (<i>n</i> = 229)	0.898	0.890	0.894	0.885	0.888	0.887	0.904	0.902	0.903	0.903	0.895	0.899
<i>R</i> ₃ (<i>n</i> = 569)	0.910	0.902	0.906	0.906	0.905	0.905	0.912	0.920	0.916	0.905	0.903	0.904
<i>R</i> ₄ (<i>n</i> = 403)	0.889	0.891	0.890	0.885	0.890	0.887	0.897	0.901	0.899	0.885	0.886	0.885
<i>R</i> ₅ (<i>n</i> = 531)	0.893	0.883	0.888	0.890	0.884	0.887	0.902	0.904	0.903	0.894	0.888	0.891
<i>R</i> ₆ (<i>n</i> = 362)	0.892	0.872	0.882	0.878	0.862	0.869	0.894	0.885	0.890	0.898	0.878	0.887
<i>R</i> ₇ (<i>n</i> = 323)	0.892	0.885	0.889	0.888	0.895	0.891	0.901	0.903	0.902	0.894	0.889	0.891
<i>R</i> ₈ (<i>n</i> = 247)	0.895	0.894	0.895	0.894	0.894	0.894	0.903	0.910	0.906	0.902	0.904	0.903
<i>R</i> ₉ (<i>n</i> = 611)	0.886	0.883	0.884	0.891	0.884	0.887	0.900	0.897	0.899	0.895	0.886	0.890
<i>R</i> ₁₀ (<i>n</i> = 439)	0.912	0.913	0.913	0.886	0.895	0.890	0.910	0.910	0.910	0.909	0.901	0.905
<i>R</i> ₁₁ (<i>n</i> = 92)	0.890	0.887	0.889	0.887	0.888	0.887	0.895	0.896	0.896	0.902	0.894	0.898
<i>R</i> ₁₂ (<i>n</i> = 184)	0.897	0.879	0.887	0.895	0.876	0.884	0.894	0.893	0.893	0.905	0.888	0.896
<i>R</i> ₁₃ (<i>n</i> = 60)	0.954	0.964	0.959	0.939	0.953	0.945	0.952	0.966	0.959	0.949	0.960	0.954

Table B2c
De-identification performance (in terms of Precision, Recall, and F-measure) based on random clusters

Train $\Rightarrow$ $\Downarrow$ Test	R_9			R_{10}			R_{11}			R_{12}			R_{13}		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
$R_1 (n = 547)$	0.888	0.882	0.885	0.869	0.842	0.855	0.866	0.852	0.859	0.853	0.830	0.841	0.850	0.741	0.791
$R_2 (n = 229)$	0.900	0.896	0.898	0.886	0.861	0.873	0.885	0.872	0.878	0.867	0.845	0.856	0.862	0.762	0.809
$R_3 (n = 569)$	0.908	0.904	0.906	0.895	0.875	0.885	0.888	0.881	0.884	0.879	0.846	0.862	0.867	0.764	0.810
$R_4 (n = 403)$	0.887	0.888	0.887	0.873	0.855	0.863	0.871	0.859	0.865	0.830	0.836	0.832	0.854	0.743	0.793
$R_5 (n = 531)$	0.888	0.887	0.887	0.879	0.853	0.866	0.880	0.861	0.870	0.860	0.822	0.840	0.856	0.756	0.801
$R_6 (n = 362)$	0.893	0.878	0.886	0.878	0.837	0.857	0.874	0.841	0.857	0.840	0.792	0.815	0.859	0.728	0.786
$R_7 (n = 323)$	0.892	0.892	0.892	0.873	0.843	0.857	0.870	0.852	0.861	0.867	0.827	0.846	0.862	0.756	0.805
$R_8 (n = 247)$	0.901	0.899	0.900	0.887	0.857	0.872	0.877	0.864	0.871	0.871	0.833	0.851	0.868	0.759	0.809
$R_9 (n = 611)$	0.891	0.888	0.890	0.884	0.849	0.866	0.883	0.856	0.869	0.865	0.825	0.844	0.873	0.743	0.803
$R_{10} (n = 439)$	0.899	0.898	0.898	0.872	0.844	0.858	0.889	0.883	0.886	0.864	0.824	0.843	0.864	0.744	0.799
$R_{11} (n = 92)$	0.892	0.891	0.891	0.879	0.856	0.867	0.890	0.872	0.880	0.867	0.848	0.857	0.855	0.724	0.783
$R_{12} (n = 184)$	0.896	0.891	0.893	0.875	0.841	0.857	0.887	0.865	0.876	0.855	0.819	0.836	0.855	0.718	0.775
$R_{13} (n = 60)$	0.944	0.969	0.956	0.926	0.926	0.926	0.943	0.952	0.947	0.913	0.878	0.894	0.931	0.818	0.869

Appendix C

In this appendix, we provide the scalability results for large random mixture clusters (Fig. C1) and complexity measure-based clusters (Fig. C2). For each subplot, the x -axis represents the size of subset invoked for trained, while the y -axis corresponds to the F -measure of the result.

Fig. C1. Scalability analysis on large mixed random subsets

Fig. C2. Scalability analysis on stylometric clusters

References

1. Jha AK, Burke MF, DesRoches C, Joshi MS, Kralovec PD, Campbell EG, et al. Progress toward meaningful use: hospitals' adoption of electronic health records. *Am J Manag Care*. 2011; 17(12 Spec No.):SP117–SP124. Epub 2012/06/27. [PubMed: 22216770]
2. 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*.

- 2013; 20(1):84–94. <http://dx.doi.org/10.1136/amiajnl-2012-001012> (Epub 2012/08/04). [PubMed: 22859645]
3. National Research Council (US). Committee on Engaging the Computer Science Research Community in Health Care Informatics. In: Stead, WW.; Lin, HS., editors. Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions. National Academy of Sciences; Washington, DC: 2009.
 4. Van Vleck TT, Wilcox A, Stetson PD, Johnson SB, Elhadad N. Content and structure of clinical problem lists: a corpus analysis. *AMIA Annu Symp Proc.* 2008;753–757. Epub 2008/11/13. [PubMed: 18999284]
 5. Johnson SB, Bakken S, Dine D, Hyun S, Mendonca E, Morrison F, et al. An electronic health record based on structured narrative. *J Am Med Inform Assoc.* 2008; 15(1):54–64. <http://dx.doi.org/10.1197/jamia.M2131> (Epub 2007/10/20). [PubMed: 17947628]
 6. Zhou, X.; Zheng, K.; Ackerman, M.; Hanauer, D. Cooperative documentation: the patient problem list as a nexus in electronic health records. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*; 2012; p. 911–920.21453402012
 7. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform.* 2008;128–144. Epub 2008/07/30. [PubMed: 18660887]
 8. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc.* 2011; 18(2):181–186. <http://dx.doi.org/10.1136/jamia.2010.007237> (Epub 2011/01/15). [PubMed: 21233086]
 9. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform.* 2009; 42(5):760–772. <http://dx.doi.org/10.1016/j.jbi.2009.08.007>. [PubMed: 19683066]
 10. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med.* 2011; 3(79):79re1. <http://dx.doi.org/10.1126/scitranslmed.3001807>. [PubMed: 21508311]
 11. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, Waudby C, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *J Am Med Inform Assoc.* 2012; 19(2):225–234. <http://dx.doi.org/10.1136/amiajnl-2011-000456> (Epub 2012/02/10). [PubMed: 22319176]
 12. Roberts A, Gaizauskas R, Hepple M, Guo Y. Mining clinical relationships from patient narratives. *BMC Bioinformatics.* 2008; 9(Suppl. 11):S3. <http://dx.doi.org/10.1186/1471-2105-9-s11-s3> (Epub 2008/12/17). [PubMed: 19025689]
 13. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc.* 2011; 18(5):540–543. <http://dx.doi.org/10.1136/amiajnl-2011-000465> (Epub 2011/08/19). [PubMed: 21846785]
 14. Standards for Privacy of Individually Identifiable Health Information. 2000.
 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. 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. <http://dx.doi.org/10.1186/1471-2288-10-70>. [PubMed: 20678228]
 17. 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. <http://dx.doi.org/10.1016/j.ijmedinf.2010.09.007> (Epub 2010/10/19). [PubMed: 20951082]
 18. Taira RK, Bui AA, Kangaroo H. Identification of patient name references within medical documents using semantic selectional restrictions. *Proc AMIA Symp.* 2002;757–761. Epub 2002/12/05. [PubMed: 12463926]

19. Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymisation of medical records using an iterative machine learning framework. *J Am Med Inform Assoc.* 2007; 14(5):574–580. [PubMed: 17823086]
20. Uzuner Ö, Sibanda TC, Luo Y, Szolovits P. A de-identifier for medical discharge summaries. *Artif Intell Med.* 2008; 42(1):13–35. <http://dx.doi.org/10.1016/j.artmed.2007.10.001>. [PubMed: 18053696]
21. 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. [PubMed: 17600096]
22. Gardner J, Xiong L. An integrated framework for de-identifying unstructured medical data. *Data Knowl Eng.* 2009; 68(12):1441–1451. <http://dx.doi.org/10.1016/j.datak.2009.07.006>.
23. Lafferty, JD.; McCallum, A.; Pereira, FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*; 2001; p. 282–289.6558132001
24. Sha, F.; Pereira, F. Shallow parsing with conditional random fields. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*; 2003; p. 134–141.10734732003
25. Settles, B. Biomedical named entity recognition using conditional random fields and rich feature sets. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*; 2004; p. 104–107.15676182004
26. McCallum, AK. MALLETT: A Machine Learning for Language Toolkit. 2002.
27. Ferrandez O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. *J Am Med Inform Assoc.* 2013; 20(1):77–83. <http://dx.doi.org/10.1136/amiajnl-2012-001020> (Epub 2012/09/06). [PubMed: 22947391]
28. Finkel, JR.; Grenager, T.; Manning, C. Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*; 2005; p. 363–370.12198852005
29. Wellner B. *Sequence Models and Ranking Methods for Discourse Parsing*. Citeseer. 2009
30. Holmes D. Authorship attribution. *Comput Humanit.* 1994; 28(2):87–106. <http://dx.doi.org/10.1007/BF01830689>.
31. Stamataatos E. A survey of modern authorship attribution methods. *J Am Soc Inf Sci Technol.* 1998; 60(3):1–28.
32. Koppel M, Schler J. Computational methods in authorship attribution. *J Am Soc Inf Sci Technol.* 2009; 60(1):9–26. doi:10./asi.1002.
33. Klare, GR. *The Measurement of Readability*. Ames, Iowa State University Press; Ames: 1963.
34. Klare GR. Assessing readability. *Read Res Q.* 1974; 10(1):62–102. <http://dx.doi.org/10.2307/86.7470>.
35. DuBay, WH. *The Principles of Readability*, 2004. Impact Information; Costa Mesa: 2008. doi: 10.1.1.91.4042
36. Flesch R. A new readability yardstick. *J Appl Psychol.* 1948; 32(3):221–233. Epub 1948/06/01. [PubMed: 18867058]
37. Aston, G.; Bernardini, S.; Stewart, D. *Corpora and Language Learners*. John Benjamins Publishing Company; 2004.
38. Tweedie FJ, Baayen RH. How variable may a constant be? Measures of lexical richness in perspective. *Comput Humanit.* 1998; 32(5):323–352.
39. Giuse DA. Supporting communication in an integrated patient record system. *AMIA Annu Symp Proc.* 2003:1065. Epub 2004/01/20. [PubMed: 14728568]
40. Gupta D, Saul M, Gilbertson J. Evaluation of a de-identification software engine: progress towards sharing clinical documents and pathology reports. *Am J Clin Pathol.* 2004; 121(2):176–186. [PubMed: 14983930]
41. 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. <http://dx.doi.org/10.1038/clpt.2008.89> (Epub 2008/05/27). [PubMed: 18500243]
42. 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.* 2013; 20(2):342–348. <http://dx.doi.org/10.1136/amiainl-2012-001034> (Epub 2012/07/10). [PubMed: 22771529]
 43. Michalke M. koRpus: An R Package for Text Analysis. 2012
 44. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Annu Symp.* 2001:17–21.
 45. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004; 32(Database issue):D267–D270. <http://dx.doi.org/10.1093/nar/gkh061>. [PubMed: 14681409]
 46. Lehner F. Quality control in software documentation based on measurement of text comprehension and text comprehensibility. *Inf Process Manage.* 1993; 29(5):551–568. [http://dx.doi.org/10.1016/0306-4573\(93\)90079-s](http://dx.doi.org/10.1016/0306-4573(93)90079-s).
 47. McCarthy P, Jarvis S. MTLT, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behav Res Methods.* 2010; 42(2):381–392. <http://dx.doi.org/10.3758/BRM.42.2.381>. [PubMed: 20479170]
 48. Anderson J. Lix and Rix: variations on a little-known readability index. *J Read.* 1983; 26(6):490–496.
 49. Deng, H.; Runger, G. Feature selection via regularized trees. *Proceedings of the International Joint Conference on Neural Networks*; 2012;
 50. Mojena R. Hierarchical grouping methods and stopping rules: an evaluation. *Comput J.* 1977; 20(4):359–363.
 51. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. *Bioinformatics.* 2005; 21(15):3201–3212. <http://dx.doi.org/10.1093/bioinformatics/bti517> (Epub 2005/05/26). [PubMed: 15914541]
 52. 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. [PubMed: 17600094]

Summary points

What was already known on the topic

- Electronic health record (EHR) systems have been widely adopted by medical institutions and benefited the quality of healthcare. Healthcare organizations still find it challenging to share clinical narratives, because patients' privacy could be subject to the risk of leaking.
- There are various software tools based on machine learning approaches for de-identification, within which CRFs have been adopted as an effective strategy. For such kind of tools, previous studies showed that training and testing on clinical narratives of the same document type yield the best performance.
- Writing styles are author-specific literary patterns, which can be characterized and assessed quantitatively using stylometric features.

What this study added to our knowledge

- Our experiments results confirm the hypothesis that training and testing on the same stylometric cluster performs better than testing on other clusters.
- Stylometric de-identification models yield better de-identification performances than random-documents-trained models.
- If document typing is not available, clustering clinical documents by their writing complexity could yield similar or in some cases more effective de-identification models.

Fig. 1. Framework for building, training, and testing de-identification models based on complexity measures and random processes

De-identified Record	Resynthesized Record
<p>**NAME[ZZZ, YYY M]</p> <p>**DATE[Sep 08 2010] 13:18: **NAME[XXX WWW] (school nurse) calling needing to clear up some things like when pt has seizures at school when pt can be sent home. Pls call 615-**PHONE .</p>	<p>Martin, Jessie C.</p> <p>Dec 15 2010 13:18: Ashley Johnson (school nurse) calling needing to clear up some things like when pt has seizures at school when pt can be sent home. Pls call 615-331-7755.</p>

Fig. 2. An example of a de-identified and resynthesized clinical narrative

Fig. 3. Framework for building, training, and testing de-identification models based on VUMC type designation

Fig. 4. VUMC document type distribution for stylometric clusters

Fig. 5. Average F -measure (± 1 standard deviation) of de-identification models as a function of the training (subset) set size for (left) large random mixture clusters, and (right) stylometric clusters

Table 1
VUMC EMR document types in the study and corresponding performances of de-identification training models

Document type	Documents	Average number of PHI instances	Precision	Recall	F-measure
History & Physical (HP)	550	23.8	0.882	0.897	0.889
General Report (REP)	550	24.64	0.941	0.922	0.931
Discharge Summary (DS)	550	47.29	0.937	0.941	0.939
Radiology Note (RAD)	547	5.64	0.954	0.914	0.933
Pathology Note (PATH)	550	16.93	0.913	0.910	0.912
Family History (FH)	200	1.12	0.949	0.768	0.842
Clinical Communication (CC)	550	23.67	0.949	0.979	0.964
Letters (LET)	550	44.2	0.887	0.905	0.896
Rehabilitation Note (REHAB)	550	44.2	0.888	0.905	0.897
		Average	0.922	0.904	0.911
		St. Dev.	0.030	0.057	0.035

Table 2
Formulas and types of complexity feature and their associated formulaic representation [33,34,38,43,46–48]

Name	Formula	Type	Dataset
Danielson-Bryan	$DB = 1.0364 \text{ CPSP} + 0.194 \text{ CPSt} - 0.6059$ CPSP, characters per space; CPSt, characters per sentence	Readability	Vanderbilt, i2b2
Dickes-Steiwert Shortcut	$V^* = 235.95993 - (\text{Var}2 * 73.021) - (\text{Var}1 * 12.56438) - (\text{Var}3 * 50.03293)$ $\text{Var}1 = \ln(\text{number of words/number of sentences}) + 1$ $\text{Var}2 = \ln((\text{number of letters/number of words}) + 1)$ $\text{Var}3 = \text{TTR (Type-Token-Ratio)}$	Readability	Vanderbilt
Flesch Reading Ease	$RE = 206.835 - 0.846wl - 1.015sl$ wl, number of syllables per 100 words; sl, average number of words per sentence	Readability	Vanderbilt
Gunning FOG Index	$FOG = 0.4 (\text{average sentence length} + \text{percentage of words of 3 or more syllables})$	Readability	Vanderbilt
Linsear Write Index	$\text{Result} = (\text{number of easy words}) + (3 * \text{number of hard words})/\text{number of sentences}$ If result is >20, divide by 2 for grade level. If number is <20, subtract 2, then divide by 2 for grade level. Easy words: two syllables or less Hard words: three syllables or more	Readability	Vanderbilt
Carroll's corrected TTR	$CTTTR(N) = \frac{V(N)}{2\sqrt{N}}$	Lexical Richness	Vanderbilt
Dugast's Uber Index	$U = \frac{\log^2 N}{\log N - \log V(N)}$	Lexical Richness	Vanderbilt
Clinical Term Ratio	$CTR = \frac{CTN}{N}$ CTN: clinical terms number	Clinical Terms Richness	Vanderbilt, i2b2
Coleman-Liau Index	First estimates cloze percentage, then calculates grade equivalent: $CL_{ECP} = 141.8401 - 0.214590 \times \frac{100 \times \text{Number of clozes}}{\text{Number of words}} + 1.079812 \times \frac{100 \times \text{number of sentences}}{\text{number of words}}$ $CL_{grade} = -27.4004 \times \frac{CL_{ECP}}{100} + 23.06395$	Readability	i2b2
Flesch-Kincaid Grade Level	$FK_{grade} = 0.39 \times \frac{\text{Number of words}}{\text{number of sentences}} + 11.8 \times \frac{\text{Number of syllables}}{\text{number of words}} - 15.59$	Readability	i2b2

Name	Formula	Type	Dataset
FORCAST Readability Formula	$\text{FORCAST} = 20 - \frac{\text{Number of 1 syllable words} \times (150 / \text{Number of words})}{10}$	Readability	i2b2
Lasbarhetsindex (LIX)	$\text{LIX} = \frac{\text{Number of words}}{\text{Number of sentences}} + \frac{\text{Number of words longer than 6 characters} \times 100}{\text{Number of words}}$	Readability	i2b2
Kuntzsch's Text-Redundanz-Index (TRI)	$\text{TRI} = (0.449 \times \text{Number of 1 syllable words}) - (2.467 \times \text{Number of punctuation marks}) - (0.937 \times \text{Number of foreign words}) - 14.417$	Readability	i2b2
Yule's K (K.Id)	$K = 10^4 \times \frac{(\sum_{i=1}^{V(N)} V(i, N)(1/N)^2) - \text{Number of tokens}}{(\text{Number of tokens})^2}$ <p>$V(i, N)$: the numbers of word types occurring i times in a sample of length N</p>	Lexical Richness	i2b2
Measure of Textual Lexical Diversity (MTLD)	MTLD = The mean length of sequential word strings in a text that maintain a given TTR value	Lexical Richness	i2b2

Table 3
VUMC de-identification performance (in terms of *F*-measure) based on clusters derived from the complexity measures and random process

Test cluster	Min	Max	Self	Min cluster	Max cluster (if not self)
<i>Sylometric clusters</i>					
<i>C</i> ₁ (<i>n</i> = 547)	0.172	0.921	0.921	<i>C</i> ₁₃	–
<i>C</i> ₂ (<i>n</i> = 229)	0.157	0.881	0.851	<i>C</i> ₁₃	<i>C</i> ₃
<i>C</i> ₃ (<i>n</i> = 569)	0.198	0.944	0.944	<i>C</i> ₁₃	–
<i>C</i> ₄ (<i>n</i> = 403)	0.176	0.872	0.872	<i>C</i> ₁₃	–
<i>C</i> ₅ (<i>n</i> = 531)	0.15	0.925	0.915	<i>C</i> ₁₃	<i>C</i> ₆
<i>C</i> ₆ (<i>n</i> = 362)	0.13	0.884	0.878	<i>C</i> ₁₃	<i>C</i> ₁₀
<i>C</i> ₇ (<i>n</i> = 323)	0.169	0.941	0.922	<i>C</i> ₁₃	<i>C</i> ₄
<i>C</i> ₈ (<i>n</i> = 247)	0.106	0.897	0.896	<i>C</i> ₁₃	<i>C</i> ₆
<i>C</i> ₉ (<i>n</i> = 611)	0.125	0.906	0.906	<i>C</i> ₁₃	–
<i>C</i> ₁₀ (<i>n</i> = 439)	0.148	0.889	0.889	<i>C</i> ₁₃	–
<i>C</i> ₁₁ (<i>n</i> = 92)	0.492	0.993	0.96	<i>C</i> ₁₃	<i>C</i> ₄
<i>C</i> ₁₂ (<i>n</i> = 184)	0.246	0.977	0.977	<i>C</i> ₁₃	–
<i>C</i> ₁₃ (<i>n</i> = 60)	0.101	0.99	0.99	<i>C</i> ₁₂	–
<i>Random clusters</i>					
<i>R</i> ₁ (<i>n</i> = 547)	0.791	0.894	0.891	<i>R</i> ₁₃	<i>R</i> ₇
<i>R</i> ₂ (<i>n</i> = 229)	0.809	0.903	0.886	<i>R</i> ₁₃	<i>R</i> ₇
<i>R</i> ₃ (<i>n</i> = 569)	0.81	0.916	0.896	<i>R</i> ₁₃	<i>R</i> ₇
<i>R</i> ₄ (<i>n</i> = 403)	0.793	0.899	0.892	<i>R</i> ₁₃	<i>R</i> ₇
<i>R</i> ₅ (<i>n</i> = 531)	0.801	0.903	0.888	<i>R</i> ₁₃	<i>R</i> ₇
<i>R</i> ₆ (<i>n</i> = 362)	0.786	0.89	0.869	<i>R</i> ₁₃	<i>R</i> ₇
<i>R</i> ₇ (<i>n</i> = 323)	0.805	0.902	0.902	<i>R</i> ₁₃	–
<i>R</i> ₈ (<i>n</i> = 247)	0.809	0.906	0.903	<i>R</i> ₁₃	<i>R</i> ₇
<i>R</i> ₉ (<i>n</i> = 611)	0.803	0.899	0.89	<i>R</i> ₁₃	<i>R</i> ₇
<i>R</i> ₁₀ (<i>n</i> = 439)	0.799	0.913	0.858	<i>R</i> ₁₃	<i>R</i> ₅
<i>R</i> ₁₁ (<i>n</i> = 92)	0.783	0.9	0.88	<i>R</i> ₁₃	<i>R</i> ₂

Test cluster	Min	Max	Self	Min cluster	Max cluster (if not self)
R_{12} ($n = 184$)	0.775	0.896	0.836	R_{13}	R_8
R_{13} ($n = 60$)	0.869	0.959	0.869	R_{13}	R_5

Table 4
VUMC de-identification performance, in terms of F -measure, for cross-cluster experiments (training cluster $\Rightarrow$ test cluster)

Clusters compared	$C \Rightarrow C$	$R \Rightarrow C$
C_1 vs R_1	0.866	0.914
C_2 vs R_2	0.756	0.893
C_3 vs R_3	0.835	0.937
C_4 vs R_4	0.884	0.858
C_5 vs R_5	0.844	0.929
C_6 vs R_6	0.877	0.872
C_7 vs R_7	0.800	0.936
C_8 vs R_8	0.831	0.863
C_9 vs R_9	0.841	0.881
C_{10} vs R_{10}	0.860	0.790
C_{11} vs R_{11}	0.434	0.967
C_{12} vs R_{12}	0.746	0.835
C_{13} vs R_{13}	0.172	0.551
Average	0.750	0.864
St. Dev.	0.210	0.105

Table 5
Number of PHI types and instances in each of the stylometric clusters in the VUMC corpus

Cluster	Number of PHI Types in training	Average number of PHI instances in training
C ₁ (n = 547)	13	12,912
C ₂ (n = 229)	9	1036
C ₃ (n = 569)	14	17,303
C ₄ (n = 403)	14	9688
C ₅ (n = 531)	13	6207
C ₆ (n = 362)	13	8613
C ₇ (n = 323)	11	3974
C ₈ (n = 247)	11	7944
C ₉ (n = 611)	13	9145
C ₁₀ (n = 439)	12	17,455
C ₁₁ (n = 92)	4	281.7
C ₁₂ (n = 184)	10	3648.6
C ₁₃ (n = 60)	3	3231
Cluster (i2b2 dataset)		
IC ₁ (n = 413)	8	24,259.5
IC ₂ (n = 476)	8	23,817.6

Table 6
i2b2 de-identification performance (in terms of Precision, Recall, and *F*-measure) based on complexity clusters and random clusters

Train $\Rightarrow$ Test	IC ₁			IC ₂		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
IC ₁ (<i>n</i> = 413)	0.964	0.955	0.959	0.963	0.948	0.955
IC ₂ (<i>n</i> = 476)	0.961	0.936	0.948	0.974	0.970	0.972
Train $\Rightarrow$ Test	IR ₁			IR ₂		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
IR ₁ (<i>n</i> = 413)	0.970	0.960	0.965	0.966	0.962	0.964
IR ₂ (<i>n</i> = 476)	0.969	0.960	0.965	0.967	0.962	0.965