Evaluation Measures for Detection of Personal Health Information

Marina Sokolova
Faculty of Medicine,
University of Ottawa
and
Electronic Health Information Lab,
CHEO Research Institute
sokolova@uottawa.ca

Abstract

Texts containing personal health information reveal enough data for a third party to be able to identify an individual and his health condition. Detection of personal health information in electronic health records is an essential part of record de-identification. Performance evaluation in use today focuses on method's ability to identify whether a word reveals personal health information or not. In this study, we propose and show that the multi-label classification measures better serve the final goal of the record de-identification.

1 Introduction

Removing personal data in documents with sensitive contents aims to protect privacy of an individual from a third party and is called *de-identification* process. De-identification of electronic health records (EHR) became an important task of applied Health Informatics (Uzuner et al., 2007; Yeniterzi et al., 2010). Properly de-identified EHR, if revealed to a third party, will not identify the patient and his health conditions.

De-identification can be viewed as personal health information (PHI) detection, followed by alternation of the retrieved information (Danezis and Gurses, 2010). The first phase, PHI detection, uses Supervised Machine Learning, Natural Language Processing and Information Extraction techniques (Meystre et al., 2010). Name, date of birth, address, health insurance number are examples of PHI that should be detected:
[Name], [age], was admitted to the [Hospital] with chest pain and respiratory insufficiency. She appeared to have pneumonia ...

The present paper focuses on evaluation practices of PHI detection. Ordinary, the quality of PHI detection is measured in counts that record correctly and incorrectly recognized PHI words and word combinations. Table 1 presents a confusion matrix for binary classification; *tp* are true PHI, *fp* – false PHI, *fn* – false non-PHI, and *tn* – true non-PHI counts. *Accuracy*, *Precision*, *Recall*, *Fscore* are used to assess PHI detection (Yeniterzi et al., 2010). It is a common practice to evaluate detection as a binary classification of word categories (e.g., accuracy of name classification in a set of EHR).

Label \ Recognized	PHI	non-PHI
PHI	*tp*	*fn*
non-PHI	*fp*	*tn*

Table 1: A confusion matrix for binary PHI classification of words.

In this study, we argue that treating PHI detection as binary word classification does not fully meet the needs of the de-identification process. We instead formulate the PHI detection as a multi-label document classification, where performance is assessed through per-document multi-class classification. We propose that the multi-label document classification better serves the final goal of the EHR de-identification. We present a case study where *Exact Match Ratio, Labelling Fscore*, *Hamming Loss, One-error* are used to assess the PHI detection results.

2 Personal Health Information

A high demand in exchange and publishing of electronic health records promoted legislative actions of the patient privacy protection. In Ontario, Canada, the Personal Health Information Protection Act (PHIPA) protects the confidentiality of personal health information and the privacy of individuals with respect to that information, while facilitating the effective provision of health care [1].

[1] http://www.health.gov.on.ca/english/providers/legislation

Proceedings of the Workshop on Biomedical Natural Language Processing, pages 19–26,
Hissar, Bulgaria, 15 September 2011.

1. Names
2. All geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code
3. Dates (other than year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89
4. Phone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social Security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. Web Uniform Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger, retinal and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code

Table 2: Health information protected by the Health Insurance Portability and Accountability Act (US).

Similar protection acts have been enabled: United States (US) has the Health Insurance Portability and Accountability Act [2], often known as HIPAA, European Union (EU) - Directive 95/46/EC, or, Data Protection Directive, although some details vary. Table 2 lists categories of personal health information which are protected by HIPAA.

Responsibility to protect patient's privacy promoted development of tools which de-identify electronic health records (EHR) (Morrison et al., 2009; Tu et al., 2010; Uzuner et al., 2007). First large-scale testing of de-identification tools showed that some of the protected categories do not appear in EHR (Uzuner et al., 2007). The absent categories included vehicle and device serial numbers, account numbers, internet protocol, URLs, and email. At the same time, references to health care providers (e.g., hospital, clinic) and professionals (e.g., doctors, nurses) frequently appeared and had been shown to reveal patient's health information. Table 3 summarizes the empirical evidence.

The de-identification systems usually benefit from the use of machine learning algorithms and text analysis methods (Meystre et al., 2010).

[2]http://www.hhs.gov/ocr/privacy/index.html

1. Age	4 Doctor	7 Location
2. Address	5 Hospital	8 Patient
3. Dates	6 ID	9 Phone

Table 3: PHI categories prevalent in EHR de-identification.

In practice, the PHI detection tools are usually trained and tested on same type of documents and/or documents originated from the same health care provider. They require a substantial amount of stored labeled training data and consume considerable time for its processing. We summarize relevant characteristics of the current PHI detection tools as follows:

- The goal of PHI detection is to detect personally identifiable information (e.g., name, address, age-identifying date).
- PHI detection applies to documents which are guaranteed to contain patient's health information (e.g., EHR).

- A common detection task is to identify whether a word bears PHI or not; for example, a phone number is PHI. Significant work was done to detect PHI indicators according to the HIPAA directives: detect and eliminate age-defining dates, postal codes, telephone numbers, social insurance numbers, etc.

3 Common Measures for PHI Detection

Currently, PHI detection methods are evaluated through their ability to correctly identify a PHI word category (e.g., *John* should be marked as a name) (Uzuner et al., 2007; Meystre et al., 2010; Morrison et al., 2009). This is done through assigning a word into two categories (i.e., binary classification), with more emphasis put on a correct labeling of PHI words.

Binary classification performance is the most general way of comparing the detection methods. It does not favour any particular application. The method's performance is assessed on all the input texts (e.g., correctly classified names in all the input discharge summaries). Introductions of new methods usually do not provide a detailed analysis of per-document detection results (Aberdeen et al., 2010; Gardner et al., 2010; Tu et al., 2010; Yeniterzi et al., 2010).

Focus on one class prevails in text classification, information extraction, natural language processing and bioinformatics. In those applications, the number of examples belonging to one class is often substantially lower than the overall number of examples. The same condition holds for the ratio of PHI words to all the words in electronic health records.

The PHI detection evaluation goes as follows: within a set of PHI categories there is a category of special interest (e.g., names). This category is designated as a *positive* class. The negative class is either *all other words* or *another PHI category*. The measures of choice calculated on the positive class are:

$$Precision = \frac{tp}{tp + fp} \quad (1)$$

$$Recall = \frac{tp}{tp + fn} \quad (2)$$

$$Fscore = \frac{(\beta^2 + 1)tp}{(\beta^2 + 1)tp + \beta^2 fn + fp} \quad (3)$$

All the three measures concentrate on the positive class (e.g., names):

Recall is a function of its correctly classified examples *tp* (e.g., *John* is classified as a name) and its misclassified examples *fn* (e.g., *John* is classified otherwise).

Precision is a function of *tp* and examples misclassified as positives (*fp*) (e.g., *Table* is classified as a name).

Fscore usually balances *Precision* and *Recall* with $\beta = 1$.

Accuracy does not distinguish between the number of correct labels of PHI and non-PHI classes. However, it approximates an over-all probability of correct classification (e.g., *John* classified as a name and *Table* is not classified as a name):

$$Accuracy = \frac{tp + tn}{tp + fn + fp + tn} \quad (4)$$

Further, we argue that performance measures other than per-class *Accuracy*, *Fscore*, *Precision*, *Recall* do apply to PHI detection and can be beneficial for EHR de-identification. Our argument focusses on the fact that the binary per-word classification leaves aside the quality of PHI detection in a document.

4 Paradox of high PHI detection and the PHI leakage risk

We claim that the currently reported PHI detection may not be sufficient to select methods which better prevent PHI leaks. We substantiate the claim by referring to the *re-identification risk*, i.e. the risk of identification of an individual from de-identified documents. The risk was actively studied for numerical data (El Emam et al., 2008).

Note that documents with undetected PHI cannot have PHI altered, thus, cannot be completely de-identified. Thus, the re-identification risk depends on accuracy of PHI detection. At the same time, accuracy computed for word categories cannot solely account for EHR de-identification.

Let's have a set of records where every record contains three names: Patient, Location and Hospital. Suppose a de-identification method correctly detects 297 PHI indicators and misses 3. Consider two outcomes:

- if the three PHI indicators are missed in the same EHR, then that EHR poses a high risk of a unique patient identification;

- however, if the three indicators are missed in three separate EHR, then the re-identification risk is substantially lower.

To have a balanced picture of the document de-identification , we add another dimension, namely, distribution of missed PHI within the de-identified documents. Based on the missed PHI, we assign a de-identified document into the following re-identification groups:

high risk : a third party can identify an individual from the document content (e.g., Patient, Location, Hospital are not detected, hence, not de-identified);

medium risk : a third party needs one or two sources of additional information to identify an individual (e.g., Hospital is detected, but Patient and Location are not detected and not de-identified);

low risk : a third party needs several additional sources to identify an individual.

Section 5 presents a case study where a highly accurate PHI detection can still leak patient's health information through not properly de-identified EHR.

5 A case study

In this section, we illustrate the paradox of reporting binary classification results for PHI detection. Our scenario presents the situation in which three detection methods achieve similar error rates on every PHI category. Nevertheless, we further show that these methods are responsible for significantly different PHI leaks.

Let's have 500 EHR, where each document reports on one patient (e.g., referral letters, discharge summaries, lab reports). Detection methods A,B,C process the documents and obtain same scores in per-word classification. For each method, *Recall* ranges from 78.5% for Locations to 98.5% for Dates; the PHI indicators are missed as follows: Age – 1, Dates – 5, Doctor – 18, Hospital – 4, ID – 5, Location – 7, Patient – 5. [3]

Thus, we can conclude that the three methods are equally strong performers in PHI detection. But will this observation always be the case? We will now show that the superiority of one methods towards other methods largely depends on the applied evaluation measures. Let's assume that the three method errors were distributed considerably differently in per-document basis:

A in each document, A has missed no more than one PHI word; thus, there were 45 documents that had 1 missed PHI;

B if B misses a patient, then it misses a doctor, date, a location and ID in the same document; for other documents, B missed no more than one PHI word; thus, there were 5 documents with 5 different PHI missed and 20 documents with 1 missed PHI;

C always misses doctor names and another PHI in the same document; other PHI words were missed "one word per document"; thus, 18 documents missed 2 different PHI and 9 documents missed 1 PHI.

We can assume that the highest risk of patient identification comes from the 5 documents with original patient and doctor names, location indicators, numerical ID and dates. The lowest identification risk comes from the documents with one un-altered PHI example.

In terms of risk levels, A leaked 45 low risk documents. B leaked 25 documents with un-detected PHI, among those 5 high-risk documents and 20 low-risk documents. C leaked 27 documents, among those – 18 medium-risk documents and 9 low-risk documents. Table 4 presents the risks associated with every method.

Method	De-identified documents		
	High risk	Medium risk	Low risk
A	–	–	45
B	5	–	20
C	–	18	9

Table 4: Risks of the de-identified documents.

Binary classification results, *Recall* and error scores, do not differentiate between the three detection methods which contribute differently to re-identification risk prevention. Consequently, they may not lead to an appropriate selection of a detection method. To find a better selection approach, we recall that the PHI detection serves as the first step of the de-identification.

[3] These results and the number of words in each PHI category would be all the information necessary to compute the binary evaluation measures.

The detection goal, therefore, is to find so much of patient information in a document that its alteration will make the patient non-identifiable. This focusses us on two characteristics of a detection method:

1. the capability with respect to PHI word categories;

2. the ability to detect PHI categories within a given document.

We will now show that the two-dimensional PHI detection evaluation can be accommodated through the multi-labelling classification setting.

6 Multi-label classification

In multi-labelled classification, the document can be classified into several of l non-overlapping categories C_i (Sokolova and Lapalme, 2009). Examples include classification of functions of yeast genes (Mewes et al., 1997), identifying scenes from image data (Li et al., 2006), text-database alignment and word alignment in machine translation (Snyder and Barzilay, 2007), etc. In text mining of medical information, multi-label classification methods can be evaluated on OHSUMED, a collection of medical references (Hersh et al., 1994). When the learning task is document topic classification, multi-labelling is often referred as multi-topic classification (e.g., classification of clinical texts based on assigned multiple disease codes ICD-9-CM (Sasaki et al., 2007)).

The quality of multi-labelling classification is assessed through either partial or complete label matching (Kazawa et al., 2005); the latter is often referred to as exact matching. For an individual PHI category C_i, the assessment is defined by tp_i, fn_i, tn_i, fp_i. The following measures evaluate the performance on per-document:

- *Exact Match Ratio* (EMR) estimates the average per-document exact classification;

- *Labelling Fscore* (LF) estimates the average per-document classification with partial matches;

- *Hamming Loss* (HL) is the average per-document per-class total error;

- *One-error* (OE) estimates the proportion of documents with the mislabeled top label.

EMR, LF, HL count correct or incorrect label identification independently of their order or rank; OE counts incorrect labelling of the top ranked label.

In the formulae below, $L_i = L_i[1], \ldots, L_i[l]$ denotes a set of class labels for x_i, $L_i[j] = 1$ if C_j is present among the labels and 0, otherwise; L_i^{class} are labels given by a method, L_i^{data} are the document labels; L_i^t is the top ranked label, I is the indicator function.

$$EMR = \frac{1}{n}\sum_{i=1}^{n} I(L_i^{class} = L_i^{data}) \quad (5)$$

$$LF = \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{l} \frac{L_i^{class}[j]L_i^{data}[j]}{(L_i^{class}[j] + L_i^{data}[j])} \quad (6)$$

$$HL = \frac{1}{nl}\sum_{i=1}^{n}\sum_{j=1}^{l} I(L_i^{class}[j] \neq L_i^{data}[j]) \quad (7)$$

$$OE = \frac{1}{n}\sum_{i=1}^{n} I(L_i^t \neq L_i^t) \quad (8)$$

For EMR and LF, higher values show a better match between input and output labels. For HL and OE, the reverse is true.

7 PHI detection as a multi-label classification

We formulate PHI detection as a multi-label classification problem, where labels represent PHI categories. In this case, a document is assigned a label if the corresponding PHI category is found in the document contents.

Let's consider three labels: a name, a location, and the "other PHI" (e.g.,a doctor, ID, age). Knowing all the three pieces of information allows for identification of an individual, i.e., represents a high re-identification risk. We assume that the input labels are set to 1 as EHR contains the patient information in all the three categories.

If a method properly detects the three categories, then the output EHR should not contain that information. Hence, all the three output labels should be 0. If a PHI word is not detected and the information leaks out, then the EHR output label for the corresponding category is 1. This implies that a poorer PHI detection is signaled by a bigger match between the input and output labels. A smaller match between the labels signals otherwise.

In terms of the measures introduced in Section 6, we interpret their values as follows:

EMR is 0 if there is no EHR with PHI missed in all the three categories; if $EMR > 0$, then there is at least one EHR with undetected PHI in all the three categories;

LF is higher when there are more EHR with several undetected PHI;

HL is lower when there are more EHR with undetected PHI;

OE is lower when more EHR contain the top PHI undetected.

We apply the multi-label measures to evaluate A,B,C performance given in Section 5. To find the top ranked label, we note that the geographic information has the biggest impact on person re-identification (Herzog et al., (2007); El Emam et al., 2008). Thus, the location is designated as the top PHI category; its detection results are used to compute OE. As an intermediate step, Table 5 reports the exact and partial label matches for the three methods. The measure values are reported in Table 6.

Method	$I(L_i^{class} = L_i^{data})$	$\sum_{j=1}^{l} \frac{L_i^{class}[j] L_i^{data}[j]}{(L_i^{class}[j] + L_i^{data}[j])}$	$\sum_{j=1}^{l} I(L_i^{class}[j] \neq L_i^{data}[j])$	$I(L_i^t \neq L_i^t)$
A	–	11.25	1455	493
B	5	7.50	1465	493
C	–	9.45	1455	493

Table 5: Counts of exact and partial label matches for A,B,C.

Method	Multi-label measures			
	EMR	*LF*	*HL*	*OE*
A	0.00	0.045	0.97	0.986
B	0.01	0.030	0.98	0.986
C	0.00	0.038	0.97	0.986

Table 6: Multi-label evaluation of methods A,B,C

EMR, the win-or-loose measure, shows that B outputs EHR with a high re-identification risk. OE shows that A,B,C output the same volume of documents in which the top ranked PHI was undetected. LF is the most discriminative measure among those applied: it marks B as the most unsafe detector, C – as a distant second, and A - as the safest detection method .

This empirical comparison shows that the A,B,C performance is not equivalent for the PHI detection, although the binary classification measures led us to believe otherwise in Section 5. In fact, the method performance can be considered significantly different if the re-identification risk is taken into account. We have shown that the multi-label classification measures account for that difference: LF differentiated between the three methods, EMR, HL differentiated between B and A, C, although they marked the performance of A,C as equivalent; OE illustrated that the three methods equally missed the location information.

In terms of the re-identification risk, EMR singles out methods with potentially higher re-identification risk, LF separates the high, medium and low risk methods; OE concentrates on the most important category; evaluation by HL is more subtle.

Efficiency of multi-labelled classification has been discussed by Kazawa et al (2005), Fujino et al (2008), Mencia et al (2010). They showed that classification costs depend on the prior knowledge about the labels (e.g., established correspondence between training and test labels) and are proportional to the number of label categories per example. We leave the analysis of efficiency of multi-labelled PHI detection for future work.

8 Related Work

Several PHI detection tools were developed and deployed to process EHR data. These tools focus on retrieval of patient's personally identifiable information, such as patient's name, address, the name and address of the health care provider or insurer. The tools de-identify clinical discharge summaries (Uzuner et al., 2008), nurse notes (Neamatullah et al., 2008), pathology reports (Beckwith et al., 2006). So far, the published work on PHI de-identification reports results in terms of binary classification. *Precision*, *Recall*, *Fscore*, *Accuracy* are reported for

PHI categories (e.g., name, address) but not for per-document performance.

A common presentation of a PHI detection method would include the use of dictionaries of local personal and geographic names. For example, in (Neamatullah et al., 2008), the authors built a system to detect PHI in nurse notes. Manual de-identification of the notes is highly accurate: the averaged manual *Precision* = 98.0%. To improve the automated de-identification, the authors use customized dictionaries of local person, geographic and health care provider names. Without the localized dictionaries, the tool's overall *Precision* is 72.5%. When the dictionaries are used, the tool's overall *Precision* is 74.9%. The tool's performance substantially varies on identification of individual categories. For person names, the use of the customized dictionaries is adverse: *Precision* = 73.1% without the dictionaries and 72.5% – with them. Location detection, in contrast, considerably improves with the use of the local dictionaries: *Precision* increases from 84.0% to 92.2% when the local information is available, *Recall* – from 37.0% to 97.0%. The use of customized dictionaries of local names, health care providers, acronyms and "do not remove" medical terms was shown to improve the PHI detection on heterogenous EHR, gathered from several regional clinics (Tu et al., 2010). The reported *Fscore* increases from 77%, without the use of customized dictionaries, to 90%, when the dictionaries are used.

Testing detection methods on altered versions of same documents is another common trend in evaluation. EHR de-identification systems are commonly trained on re-synthesized records, i.e. records where real identifiers are substituted by synthetic ones. The re-synthesis effects on personal information detection were studied in (Yeniterzi et al., 2010); the researchers used the de-identification system first introduced by (Aberdeen et al., 2010). The system's *Fscore* declined from 98.0%, when tested on re-synthesized records, to 72.8% when tested on original records. When trained on records with original PHI, the system's performance fluctuated less: *Fscore* was 96.0%, when tested on original records, and 86.2%, when tested on re-synthesized records.

The reported accuracy, however high, does not provide enough data for a thorough understanding of the PHI de-identification. We suggest to incorporate the re-identification risk in the reported results. This can be done, for example, through per-document performance evaluation.

Relations between disclosed parts of personal information are studied for social networks. In (Al-Faresi et al, 2010), the authors apply Bayesian networks to model the risk of re-identification from email and a forum post. In (Domingo-Ferrer, 2009; Domingo-Ferrer and Saygin, 2009), the authors discuss privacy risk scores where private data categories are assigned sensitivity weights. The authors do not report how the weights should be calculated or what private categories are suggested. De-identification and re-identification processes are also left out of the study scope.

9 Conclusions and Future Work

Prevention of patient PHI leaks into a public domain has traditionally been an obligation of those responsible for the safeguarding of the information. Such obligation, reinforced by legislative requirements, prompted development of PHI de-identification methods and tools. Machine Learning, Natural Language Processing and Information Extraction techniques became essential parts of the de-identification process.

In this study, we have proposed a new approach to the evaluation of PHI detection, the first part of the de-identification. Our approach focuses on the method's ability to detect the PHI information within a document. We have argued that this not yet been done in studies of PHI detection. We proposed performance measures which originate in multi-label classification: *Exact Match Ratio, Labelling F-score, Hamming Loss, One-error*. Our case study of electronic health record de-identification has presented an application which benefits from the use of these measures. We also have presented PHI detection as the multi-label classification problem.

Our future work will follow several interconnected avenues: incorporate PHI alteration in free-form texts, find new characteristics of the methods that must be evaluated, consider new measures of method performance, and search for PHI detection applications other than EHR de-identification.

Acknowledgments

This work has been supported by NSERC and CIHR operating research grants.

References

Aberdeen, J. Bayer, S., Yeniterzi, R., Wellner, B., Clark, C., Hanauer, D. , Malin, B., Hirschman, L. The MITRE Identification Scrubber Toolkit: Design, training, and assessment. *International Journal of Medical Informatics*, 2010. 79(849–859)

Al-Faresi, A., A. Alazzawe, A. Alazzawe, and Duminda Wijesekera. Risk Analysis Framework & Architecture for DLP Systems, *Proceedings of IMPPCD-2010*, p.p. 35–43, 2010. http://www.ehealthinformation.ca/documents/IMPPCD-2010.pdf

Beckwith, B., R. Mahaadevan, U. Balis, and F. Kuo, Development and evaluation of an open source software tool for de-identification of pathology reports. *BMC Medical Informatics and Decision Making*, 2006; 6:12;

Danezis, G. and S. Gurses. A critical review of 10 years of Privacy Technology. In *the Proceedings of Surveillance Cultures: A Global Surveillance Society?*, 2010.

Domingo-Ferrer, J. The Functionality-Security-Privacy Game. *Proceedings of MDAI*, 2009, 92–101.

Domingo-Ferrer, J. and Y. Saygin. Recent progress in database privacy. *Data and Knowledge Engineering*, **68** (11), 1157–1159, 2009.

El Emam,K., A. Brown, P. Abdel Malik. Evaluating Predictors of Geographic Area Population Size Cutoffs to Manage Re-identification Risk. *Journal of American Medical Informatics Association*, **16**(2), 256–266, 2008.

Fujino, A., H. Isozaki, J. Suzuki. Multi-label Text Categorization with Model Combination Based on F1-score Maximization. *Proceedings of IJCNLP*, 2008, 823–828.

Gardner, J., L. Xiong, F. Wang, A. Post, J. Saltz, T. Grandison. An Evaluation of Feature Sets and Sampling Techniques for De-identification of Medical Records. *Proceedings of IHI*, 2010, 183–190.

Hersh, W., C. Buckley, T. Leone, and D. Hickam, 1997. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-97)*, pp. 192-201

Herzog, T., F. Scheuren, and W. Winkler, *Data Quality and Record Linkage Techniques*. 2007: Springer

Kazawa, H., Izumitani, T., Taira, H., and E. Maeda, 2005. Maximal margin labeling for multi-topic text categorization. In *Advances in Neural Information Processing Systems*, Vol. 17. pp. 649–656.

Li, T., Zhang, C., and S. Zhu, 2006. Empirical studies on multi-label classification. In *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, pp. 86–92.

Mencia, E., S.-H. Park, J. Furnkranz. Efficient voting-predictionforpairwisemultilabelclassification. *Neurocomputing*, **73**, p.p. 1164–1176, Elsevier, 2010.

Mewes, H.-W., K. Albermann, K. Heumann, S. Lieb, and F. Pfeiffer, 1997. MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Research*, 25(1), p.p. 28-30.

Meystre, S., F. Friedlin, B. South, S. Shen, and M. Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol*. 2010; 10: 70

Morrison, F., et al., Repurposing the Clinical Record: Can an Existing Natural Language Processing System De-identify Clinical Notes? *Journal of American Medical Information Association*, 2009. 16(1):37-39;

Neamatullah, I., et al., Automated De-Identification of Free-Text Medical Records. *BMC Medical Informatics and Decision Making*, 2008; **8**(32);

Sasaki, Y., Rea, B., and S. Ananiadou, 2007. Multi-topic Aspects in Clinical Text Classification. In *Proceedings of the 2007 IEEE international Conference on Bioinformatics and Biomedicine*, IEEE Computer Society, pp. 62–70.

Sokolova, M. and G. Lapalme. A systematic analysis of performance measures for classification tasks, *Information Processing and Management*, **45** (4), 427–437, 2009.

Snyder, B., and R. Barzilay, 2007. Database-text Alignment via Structured Multilabel Classification, In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI-2007)*, pp. 1713–1718.

Tu, K., J. Klein-Geltink, T. Mitiku, C. Mihai and J. Martin. De-identification of primary care electronic medical records free-text data in Ontario, Canada.; *BMC Medical Informatics and Decision Making*, 2010; 10:35;

Uzuner, O., Y. Luo, and P. Szolovits, Evaluating the state-of-the-art in automatic de-indentification. Journal of the American Medical Informatics Association, 2007; 14: 550–563.

Uzuner, O., et al., A de-identifier for medical discharge summaries. *Journal of Artificial Intelligence in Medicine*, 2008; 42:13–35;

Yeniterzi, R., Aberdeen, J., Bayer, S., Wellner, B., Hirschman, L., Malin, B. Effects of personal identifier resynthesis on clinical text de-identification. *Journal of American Medical Information Association*, 2010. 17: 159–168