Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records

Emmanuel Chazard [a,*], Capucine Mouret [b], Grégoire Ficheur [a], Aurélien Schaffar [a], Jean-Baptiste Beuscart [c], Régis Beuscart [a]

[a] Department of Public Health, CHU Lille, UDSL EA 2694, Univ Lille Nord de France, F-59000 Lille, France
[b] Department of Occupational Medicine, CHU Lille, F-59000 Lille, France
[c] Department of Geriatrics, CHU Lille, UDSL EA 2694, Univ Lille Nord de France, F-59000 Lille, France

What was already known:

- Several methods exist for free-text de-identification
- Pattern matching methods require that dictionaries are already available
- Machine learning require that a corpus of manually de-identified free-text is available
- There is no freely-available method for French language

What this study added to our knowledge:

- FASDIM is a new method related to pattern matching. It brings good results in French (recall 98.1%, precision 79.6%, and F-measure 87.9%).
- The effect of de-identification can be evaluated by measuring how much of the medical content is retained after de-identification, by means of coding (e.g. ICD10, ATC, CCAM).
- FASDIM preserves 99.02% of the codes through the de-identification.

Abstract

Purpose

Medical free-text records enable to get rich information about the patients, but often need to be de-identified by removing the Protected Health Information (PHI), each time the identification of the patient is not mandatory. Pattern matching techniques require pre-defined dictionaries, and machine learning techniques require an extensive training set. Methods exist in French, but either bring weak results or are not freely available. The objective is to define and evaluate FASDIM, a Fast And Simple De-Identification Method for French medical free-text records.

Methods

FASDIM consists in removing all the words that are not present in the authorized word list, and in removing all the numbers except those that match a list of protection patterns. The corresponding lists are incremented in the course of the iterations of the method.

For the evaluation, the workload is estimated in the course of records de-identification. The efficiency of the de-identification is assessed by independent medical experts on 508 discharge letters that are randomly selected and de-identified by FASDIM. Finally, the letters are encoded after and before de-identification according to 3 terminologies (ATC, ICD10, CCAM) and the codes are compared.

Results

The construction of the list of authorized words is progressive: 12 hours for the first 7,000 letters, 16 additional hours for 20,000 additional letters. The Recall (proportion of removed protected health information, PHI) is 98.1%, the Precision (proportion of PHI within the removed token) is 79.6% and the F-measure (harmonic mean) is 87.9%. In average 30.6 terminology codes are encoded per letter, and 99.02% of those codes are preserved despite the de-identification.

Conclusion

FASDIM gets good results in French and is freely-available. It is easy to implement and does not require any predefined dictionary.

Manuscript

I. Introduction

A. A need for de-identifying discharge letters

Computerized free-text medical records are important information sources for research. In most countries, each time a patient is discharged from a healthcare facility, a discharge letter has to be written: it summarizes all the pertinent information from the reason for admission to the discharge drug treatment. Those letters are routinely produced and provide the researchers with a big amount of medical information. On the other hand, the confidentiality must imperatively be respected: as soon as a discharge letter is not used with direct benefit to the patient and if the patient doesn't need to be identified, the letter must be de-identified. The anonymization consists in removing the patients' names from the records: unfortunately, other pieces of information enable to identify the patients. The de-identification is a more exhaustive removal of the entire Protected Health Information (PHI), so that the patients cannot be identified, directly nor indirectly. In the US, privacy rules have been enacted by the Department Of Health and Human Services further to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [1]. In order to de-identify a high number of records, it is necessary to use automated methods, as manual methods require too high workload [2].

B. State of the art

Several methods exist for automated de-identification of free-text records [3], including procedures reports and discharge letters.

Pattern matching methods [4-16] consist in applying rules that enable to keep or remove some words that belong to dictionaries that have been predefined by experts or institutions. For instance, it is possible to remove all the words that belong to a list of town names, or to preserve all the words that belong to a list of medical terms (such as the Unified Medical Language System [17]). Additional rules may be used to take into account words declension and verbs conjugation. This approach requires that such lists are available. When they exist, those lists are language-dependent, and are suitable for a specific context only (e.g. town names or current family names are useless in another country).

Machine learning methods [14, 18-26] are derived from artificial intelligence. A learning phase requires that a corpus of records is previously de-identified manually by experts. Those methods are often very efficient, depending on the quality and the completeness of the learning corpus.

Whatever the method used, the de-identification is evaluated by computing three rates:

- The recall (or sensitivity or completeness, Equation 1), which is the proportion of removed token within the PHI. A high recall enables to preserve the confidentiality.

- The precision (or positive predictive value or correctness, Equation 2), which is the proportion of PHI within the removed token. A high precision enables to preserve the readability of the text.
- The F-Measure, which is the harmonic mean of the recall and the precision (Equation 3).

$$recall = R = \frac{TP}{\#identifier\,s} = \frac{TP}{TP + FN} \tag{1}$$

$$precision = P = \frac{TP}{\#removed} = \frac{TP}{TP + FP} \tag{2}$$

$$F - measure = F = (\frac{R^{-1} + P^{-1}}{2})^{-1} \tag{3}$$

Table I presents the main results obtained in the literature by the authors for medical free-text de-identification. Most of methods are developed for English language and can hardly be used for other languages. Some methods have been developed in French, but either their results are disappointing, or they are not freely available.

Table I. Results of authors for medical free-text records de-identification.

Author	Method	Language	Precision	Recall	F-measure
Aberdeen 2010 [18]	Machine learning	EN	94.3%	97.8%	96%
Aramaki 2006 [19]	Machine learning	EN	-	-	-
Beckwith 2006 [4]	Pattern matching	EN	98.3%	-	-
Deleger 2013 [25]	Machine learning	EN	92.8%	92.8%	92.8%
	Machine learning	EN	95.1%	91.9%	93.5%
	Manual	EN	93.9%	92.1%	93%
	Manual	EN	88.5%	94.6%	91.4%
Dorr 2006 [2]	Manual	EN	95.9%	99.6%	-
Ferrandez 2012 [26]	Machine learning	EN	96%	70%	74%
	Machine learning	EN	95%	76%	79%
Friedlin 2008 [7]	Pattern matching	EN	-	99.47%	-
Grouin 2009 [9]	Pattern matching	FR	92%	83%	-
		EN	23%	65%	-
Neamatullah 2008 [11]	Pattern matching	EN	75%	94%	-
Ruch 2000 [12]	Pattern matching	FR	-	99%	-
Sweeney 1996 [13]	Pattern matching	EN	-	-	-
Szarvas 2007 [22]	Machine learning	EN	-	-	96%
Taira 2002 [14]	Pattern matching & Machine learning	EN	99%	94%	-
Thomas 2002 [15]	Pattern matching	EN	-	98.7%	-
Tu 2010 [28]	Pattern matching	EN	91.3%	88.3%	90%
Uzuner 2007 [23]	Machine learning	EN	99%	97%	98%
Velupillai 2009 [16]	Pattern matching	SW	3-9%	56-76%	4-16%
Wellner 2007 [24]	Machine learning	EN	98%	96%	96%

C. Unsolved situations

Despite the good results obtained by many methods, text de-identification is still not obvious and some situations may not be addressed with current tools. We shall illustrate it through 4 use cases.

Case 1: a team has to de-identify English free-text records using pattern-matching. Some tools are freely available. However, it cannot be guaranteed that those tools could be applied in a different context without any adaptation. Indeed, pattern matching techniques rely on lists of words that are context-dependent: for instance "lime tree" should be removed in most reports as it is often part of a street name, but shouldn't be removed in an allergy-related report. Lists of town names or family names also depend on the country. Finally, misspellings are most often not taken into account by existing methods.

Case 2: a team has to de-identify English free-text records using machine learning. Here again, some tools are freely available but, in a like manner, machine learning techniques require a pre-existing corpus of de-identified records. Such corpuses are available in English [11, 36, 37], but they may be used only if the type of document to de-identify is the same as the documents of the training corpus.

Case 3: a team has to de-identify French free-text records (the problem is the same with most of non-English languages): no free and efficient method, no list of words, and no training corpus are available. Everything has to be built.

Case 4: a team has only little time (e.g. 1 man-week) to de-identify a few records (e.g. 25,000 records). Whatever the language, the context and the technique, it will probably take more time to understand, adapt, implement and execute an existing tool.

The conception of FASDIM relies on the idea that a simple de-identification technique could enable to de-identify French discharge letters with an acceptable workload, particularly when the number of records is low. The main idea is to supply the workload in the course of the method, and not before the first document can be de-identified.

D. Objectives

The first general objective of this work is to design and implement FASDIM, a Fast And Simple De-Identification Method for clinical free-text records. The second general objective is to evaluate the method.

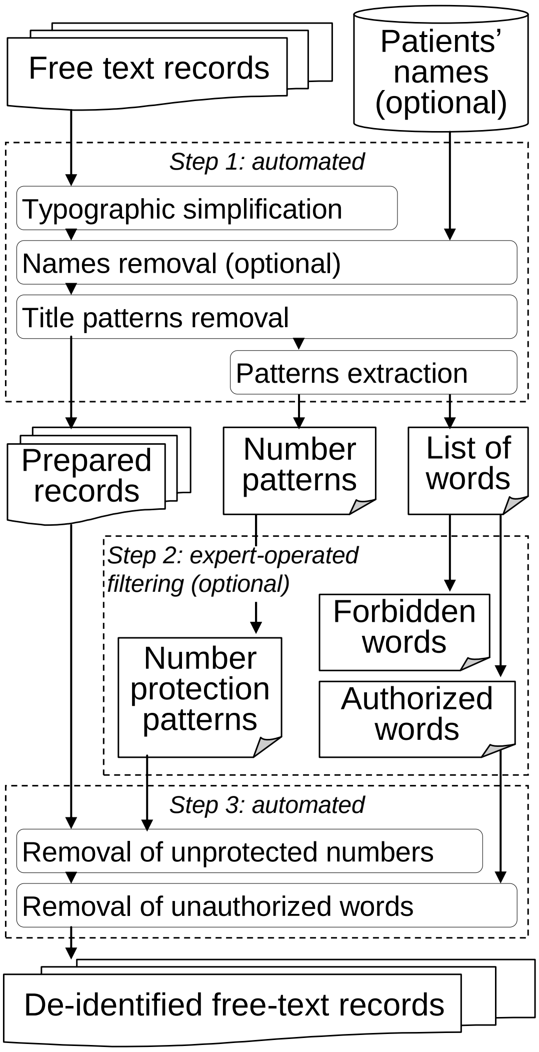
To reach the first general objective, operational objectives are (1) to design a method that reaches good results in French using completely unstructured free-text records, but (2) is as independent as possible from the language structure (i.e. for instance doesn't consider the declension of words and the conjugation of verbs) and (3) doesn't rely on any pre-existing material (list of words or corpus of de-identified documents), in order to (4) be easily and fast reproducible from scratch by any hospital or research team.

To reach the second general objective, operational objectives are (5) to objectively compute traditional evaluation metrics but also (6) to evaluate the preservation of medical information and (7) to evaluate the workload required to implement the method.

The method is implemented and evaluated in French, but the examples that are given in this paper here are translated into English.

II. Definition of FASDIM

FASDIM stands for Fast And Simple De-Identification Method. This method is composed of 3 steps (Figure 1).



Free text records
Patients' names (optional)
Step 1: automated
Typographic simplification
Names removal (optional)
Title patterns removal
Patterns extraction
Prepared records
Number patterns
List of words
Step 2: expert-operated filtering (optional)
Forbidden words
Number protection patterns
Authorized words
Step 3: automated
Removal of unprotected numbers
Removal of unauthorized words
De-identified free-text records

Figure 1. The algorithm of FASDIM consists of 3 steps. Steps 1 and 3 are fully automated. Step 2 is expert-operated but does not necessarily require to be performed for each iteration of FASDIM.

A. Step 1 (automated): preparation of the records and patterns extraction

The first step of the algorithm consists of an automated treatment of the free-text records (Figure 1). The records are loaded as simple text, and the typography is simplified: the text is lowercased, the accents are removed from the letters, and special characters are replaced by simpler characters (Figure 2).

é, è, ê, ë → e   œ → oe   ç → c

Figure 2. Examples of typographic simplification

Optionally, the first name and the last name of the patient can be available in the Hospital Information System (HIS), with a link to his or her letter(s). In that case, the first name and last name of the patient are extracted from the HIS and are removed from the corresponding letters. For that purpose, the first name and the last name are split on spaces and punctuation marks, and each token is removed from the text. This step is optional. Finally, the titles (Mr, Mrs, Dr, etc.) and the 1 or 2 following words are removed by means of 48 regular expressions. An example is displayed on Figure 3.

Regular expression: \bmr\.\s+\w+\s+\w+\b *(case insensitive)*
which means *[WB]"mr."[WS][word] [WS] [word] [WB]*
with *WB=word boundary, WS=whitespace character(s)*
as defined in Perl-compatible regular expressions
Original string: *"I have examined Mr. James Jones."*
Transformed string: *"I have examined @ @ @."*

Figure 3. Example of title pattern removal

This process enables to get "prepared records". Finally, regular expressions are applied to those records in order to extract (1) a list of all the available words, as well as their frequencies, and (2) a list of all the different patterns that involve numbers (each pattern include the word before and the word after a number), as well as their frequencies.

The results of the first step are:

- a set of prepared records, that are ready to undergo the third step
- a list of words, that will be filtered in the second step
- a list of patterns including numbers, that will be filtered in the second step

B. Step 2 (expert-operated, sporadic): filtering of the lists

The second step consists of a review and a manual filtering of the lists generated in the first step (Figure 1). It is important to understand that this step can even be performed very fast on a little set of letters, depending of the needs. In addition, it can also be performed incrementally e.g. if there is a certain number of records to de-identify each month.

Experts are asked to review the list of patterns including numbers that are discovered in the first step. The experts can review the corresponding letters stored in the database if necessary. From the list, they define a list of "number protection patterns": all the numbers that match the patterns can be kept without confidentiality threat. This is illustrated on Figure 4. Those number protection patterns mostly include prepositions, measure units, clinical parameters, galenic forms and drug names.

Example of number to delete:	*"...has been discharged the 24th January…"*
Example of number to protect:	*"...a respiratory rate of 24 breaths/minute…"*
Number protection pattern:	*\b\d+\sbreath (case insensitive)*
which means	*[WB][number][WS]"breath"*
with	*WB=word boundary, WS=whitespace character*

Figure 4. Example of number consideration

The experts are also asked to review the list of different words discovered in the first step. They can review the corresponding letters stored in the database if necessary. They filter that list in order to get a list of "authorized words": all the words of this list can be kept without confidentiality threat. The other words are put into a list of "forbidden words". That second list is not useful for the third step, but prevents from reviewing those words again during next iterations of the second step. This second step is crucial and the list of authorized words is not simply a list of common words:

- this list should include some words that are not common words, such as:
  - some misspellings, e.g. "Ferosemide" instead of "Furosemide",
  - some medical proper names, e.g. "Prader-Willi",
  - some medical abbreviations or acronyms, e.g. "HTN" for "high blood pressure".
- this list should exclude some words despite they are common words, such as:
  - words that refer to dates, e.g. "tomorrow", "Monday", "January",
  - words that refer to places, e.g. "street", "cardiology", "hospital", "town",
  - words that are frequently present in street names, town names or names of healthcare facilities, e.g. "liberty", "square", "street" or "forest".

Many choices at this step are not obvious, and those choices are probably impacted by the context of the de-identification. However the experts are asked to value confidentiality over legibility of the de-identified text.

C. Step 3 (automated): de-identification

Finally, the numbers that match the number protection patterns defined in step 2 are protected. All the numbers that are not protected are removed from the text. All the words that do not belong to the authorized list are removed from the text. At the end of the process, the text is de-identified (Figure 1).

D. Practical use of the 3 steps

Steps 1 & 3 are fully automated. Step 2 is an expert-operated filtering of lists and patterns. Contrary to classical pattern-matching techniques, the lists do not have to be written before the de-identification process: they are filtered in the course of the use of the method. If a small number of records are de-identified, the list of words and patterns is short and then can be filtered very fast. FASDIM could typically be used as follows. During the first iterations, the 3 steps are performed. During next iterations, the 2nd step can possibly be discarded. The absence of 2nd step does not threat the confidentiality: the only risk is to over-scrub, i.e. to remove too many tokens from the text. Indeed, lists of words and patterns are only used to protect some tokens of the records, and never to identify the words that should be deleted, contrary to some other pattern matching methods. However the 2nd step should be performed from time to time. Another way is to directly get the list of words and patterns from another user of the method [27].

III. Material & method of the evaluation

The FASDIM method has been first developed to meet the needs of a research project, with imposed deadlines: that explains why the numbers of records at each step are not regular. Seven successive sets of unstructured discharge letters are extracted from the HIS of a general French hospital:

- A first set of 20 records used to develop and test the method
- Successive cumulative sets of records: 7,012 then 9,503 then 16,009 then 17,812 then 23,493 letters
- Finally, from the last cumulated extraction of 28,540 records, 1,000 records that do not belong to the 23,493 first records are randomly selected to build an evaluation set, and are excluded from the learning set.

This way we obtain 6 cumulated learning sets (the latest one contains 27,540 letters) and 1 evaluation set of 1,000 records (due to time restrictions, only the first 508 of them are annotated by the experts and used for the evaluation). The names of the corresponding patients are simultaneously extracted from the HIS, with an identifier that enables to link each patient name to the corresponding letters.

A list of the categories of PHI to remove is obtained from the HIPAA [1]. That list is complemented using the names and addresses of healthcare providers as, according to several authors, they could enable to identify the patients [4, 7, 11, 12, 22, 24, 28, 29]. The list of PHI categories is presented in Table II.

Table II. The list of protected health information categories is obtained from the HIPAA. The items marked with (*) are frequently added by authors and will be used in this work.

	Protected Health Information categories
1.	Names
2.	Geographic subdivisions smaller than a State
3.	All elements of dates (except year), ages over 89
4.	Telephone 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
13.	Device identifiers and serial numbers
14.	Web Universal Resource Locators (URLs)
15.	Internet Protocol (IP) address numbers
16.	Biometric identifiers, including finger and voice prints
17.	Full face photographic images
18.	Any other unique identifying number, characteristic, or code
19.	(*) Names of health personnel, or health facilities
20.	(*) Geographic location of health facilities

The evaluation consists of 3 phases. For phases 1 & 2, the 508 unstructured discharge letters are de-identified by FASDIM, using the patients' names (Figure 5). The evaluation phases 1&2 are performed by 3 independent experts who are physicians and are aware of confidentiality rules and health terminologies. The third evaluation phase is performed by the developer of the tool. Ninety-five percent confidence intervals are provided when appropriate.

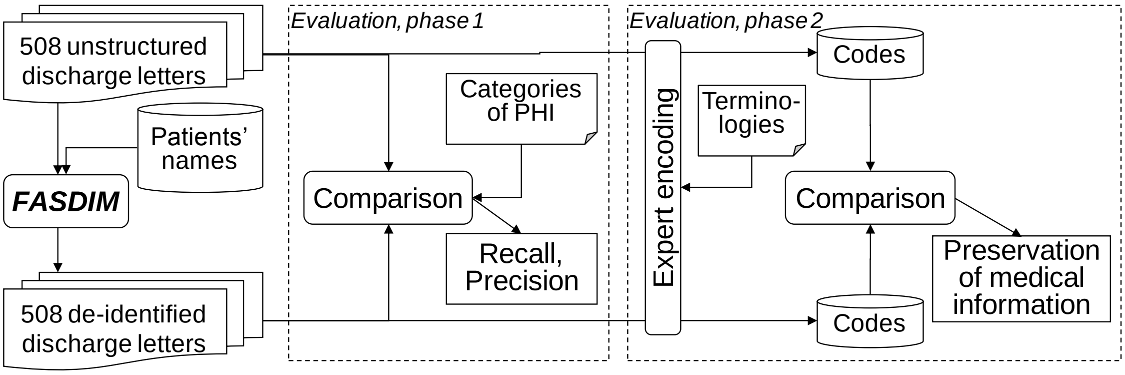


Figure 5. The first evaluation phase measures the recall and precision of the de-identification. The second evaluation phase measures the preservation of medical information according to three terminologies.

A. First evaluation phase

The 508 original discharge letters and the 508 de-identified letters are reviewed by an independent expert (Figure 5, middle). The expert is in charge of counting:

- the PHI tokens that are removed by FASDIM (true positives, TP)
- the PHI tokens that are not removed by FASDIM (false negatives, FN)
- the tokens that are removed by FASDIM but are not PHI (false positives, FP)

The counting process is strict. For instance, if an error appears several times in the same letter, it is counted several times also, which is not systematic in the literature [30]. The numbers above enable to compute the precision, the recall and the F-measure as defined in the introduction section (Equations 1-3).

B. Second evaluation phase

If a token is falsely removed (i.e. false positive), it might unequally alter the readability of the document, and in particular the medical information: indeed, the removed token could be either an insignificant word or a medical term. The second evaluation phase deals with that issue by evaluating the preservation of medical information. For that purpose (Figure 5, right), experts are asked to encode the anonymized discharge letters using several terminologies. Then, the same experts encode the original discharge letter using the same terminologies. It is chosen to encode exhaustively all the concepts (e.g. a disease and all the related symptoms that are described in the letter). This enables to compare the codes that are chosen after and before the anonymization process and thus to compute the preservation rate of the medical information (number of codes after / number of codes before). The terminologies are:

- The ATC for the drug names [31]
- The ICD10 for three categories of information [32]:
    - Diseases, symptoms and other factors (most of the codes)
    - Some abnormal laboratory results (e.g. R73.9 - hyperglycemia)
    - Some medical procedures (e.g. Z49.1 - extracorporeal dialysis)
- The CCAM, a French classification, for the medical diagnostic or therapeutic procedures [33]

C. Third evaluation phase

The aim of the third phase is to measure the cumulated time required to implement the method and to perform the second step of the method, which consists in filtering the number patterns and the list of authorized words, this second task being from far the most important one.

D. Ethics

The persons who had access to real free-text medical letters are all physicians (doctors or students) and are bound by professional secrecy. The medical letters have been handled by respecting confidentiality rules. The study is covered by the general authorization of the hospital about observational studies. All the patients of the hospital are informed that their medical data can be used for observational studies.

IV. Results of the evaluation

A. First evaluation phase

The accuracy of the de-identification is evaluated using 508 discharge letters (Table III). The recall is 98.1% [97.8% ; 98.4%], the precision is 79.6% [78.9% ; 80.3%] and the F-measure is 87.9%. Many auxiliary verbs are over-scrubbed because they stand nearby family names, but it doesn't alter the legibility of the text. If the suppression of auxiliary verbs is ignored, the precision reaches 89.2% and the F-Measure reaches 93.4%.

Table III. Results of the first evaluation phase.

Measures	Values
Number of discharge letters	508
Total number of PHI	9,914
Mean number of token per letter	510
Mean number of PHI per letter	19.5
False positives (FP)	2,537
False negatives (FN)	183
Mean number of FN per letter	0.36
Recall (R)	98.1%
Precision (P)	79.6%
F-measure (F)	87.9%

In average, 0.36 PHI token are inappropriately preserved per letter [0.318 ;0.402]. Those PHI token can be categorized as in Table IV. None of those PHIs is a patient name or a complete date.

Table IV. Description of the false negatives (PHI inappropriately preserved).

Categories of forgotten PHI token	Proportion
Partial information about a place	63.7 %
Healthcare professional's name	23 %
Patient's weight	5.5 %
Part of date or patient's age	4.4 %
Health facility name	3.3 %

B. Second evaluation phase

The preservation of medical information is evaluated through a double encoding process using 3 terminologies, before and after the de-identification. Each letter contains in average 30.6 codes from those terminologies (15,563 codes in 508 letters). Despite the de-identification, 99.02% of the codes are preserved. This rate is detailed per terminology in Table V.

Table V. Preservation rate of the medical information

Terminology	Preservation rate
CCAM - medical procedures	99.66%
ICD10 - diagnoses, symptoms & others	99.49%
ICD10 - procedures	98.92%
ICD10 - abnormal laboratory results	96.99%
ATC - drugs	98.84%
All terminologies	99.02%

C. Third evaluation phase

The time required to develop FASDIM is displayed on Figure 6. An incompressible time has been necessary to develop a simple and functional version of the program (12 hours). Then, additional time is required for each iteration (respectively 7,012, 9,503, 16,009, 17,812, 23,493 then 27,540 letters) to update the number protection patterns and mostly the list of authorized words. In summary, 28 hours are necessary to de-identify 7,000 letters or 40 hours for 27,000 letters when no pre-existing material or piece of software is available.

Indeed, additional letters bring new unlisted words (Figure 7), with an increasing proportion of misspellings [28], but the de-identification process has to take them into account. After de-identification of 27,540 discharge letters, the lists contain about 17,600 authorized words and 512 number protection patterns. Those lists are freely available on the Web [27] and will be updated so that it should require less time for another team to use the FASDIM method on French discharge letters.

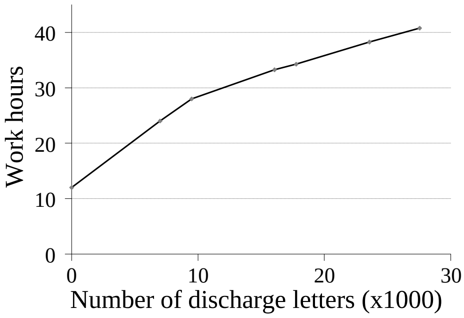


Figure 6. Workload as a function of the number of discharge letters to de-identify (including the development of the software).

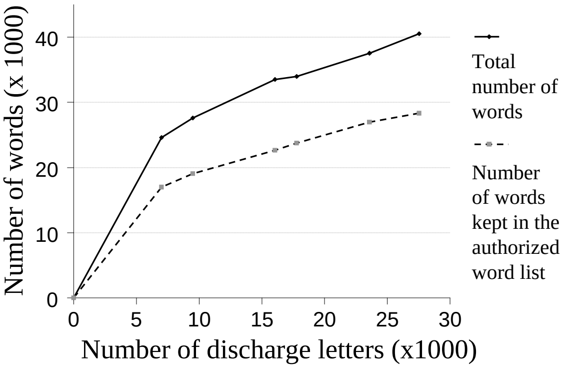


Figure 7. Number of different words as a function of the number of discharge letters.

V. Discussion

In this work, FASDIM, a Fast And Simple De-Identification Method for clinical unstructured free-text records, has been defined and evaluated in French language. The operational objectives defined in the Introduction section have been reached (Table VI).

Table VI. Summary of the operational objectives defined in the Introduction section

#	To design a method that:
1	reaches good results in French
2	is as independent as possible from the language structure
3	doesn't rely on any pre-defined corpus of words
4	can be easily reproducible by any hospital or research team
#	**To evaluate the method by:**
5	computing traditional evaluation metrics
6	evaluating the preservation of medical information
7	evaluating the workload required

Objectives 1, 5 & 6: the method reaches very good results in French. The recall is 98.1%, the precision is 79.6% and the F-measure is 87.9%. A less strict evaluation gives a precision of 89.2% and an F-Measure of 93.4%. Moreover, 99.0% of the medical information is preserved after the identification.

Objective 2: the method doesn't rely on a strong knowledge of the language. Declensions, conjugations, syntax of the sentences and semantic links within sentences are ignored. That enables to get a simple and fast to develop method; however the results remain good enough.

Objective 3: the method doesn't rely on any pre-defined corpus of words: such a corpus is defined iteratively during the second step of the method (Figure 1). That enables to get de-identified letters very early, and to require only little time when there are few letters to de-identify.

Objective 4 & 7: the method can easily be reproduced with any material by any hospital or research team and a little time is required. Indeed, objectives 2 & 3 make the method very simple to implement. As a consequence, first results can be obtained fast, as shown in the workload evaluation. In summary, it seems possible to spend 12 hours to develop the software, and then in average 1 hour of additional work for each new set of 1,000 letters. On the contrary, in traditional pattern matching methods, the lists of words and patterns must be defined before de-identifying the first letter, That evaluation enables a team who would like to use or reproduce the method to predict the required workload. It also suggests that other methods should be preferred to de-identify very large amount of letters (more than 200,000), this is probably due to the fact that FASDIM doesn't support language-dependent advanced features, such as declensions and conjugations.

The method has several advantages. In contrast to traditional pattern matching, no predefined list of words is necessary. This is an advantage in particular for countries and languages where such lists are not available, or when misspellings are frequent in the letters. In contrast to machine learning, no learning corpus of de-identified free-text records is required, and the method is simple to implement from scratch. Despite a strict evaluation process, the method provides results that are comparable to the best methods in English and French.

The main drawback of FASDIM is to require an additional work to filter the new words to increment the authorized word list. This task is a tedious work that requires an implicit knowledge about language, care and medicine, and medical context. In our study it has been performed by physicians who were allowed to read the letters. However, this task requires much less time than constructing or adapting exhaustive lists. The lists that have been used can be downloaded from the Web [27] or in the additional material. However, they should probably be adapted function of the context of use. Indeed the performance of de-identification methods depend on the nature of the documents [18], and de-identification may also be applied to non-medical records such as nursing documents [35]. For instance, we have considered “acacia” as being frequently associated with street names, but such a word shouldn't be removed from records of occupational medicine or allergy. Another example: free text written in short fields such as in electronic health records may contain more abbreviations or typos. In both examples, using the method without updating the lists of words could lead to over-scrub the text. It would alter the legibility of the documents but should not threaten patient confidentiality.

FASDIM has another drawback: it is able to delete PHIs, but it is not able to tag them or to determine their type (e.g. name, address, date, etc.). This is due to the fact that the method only identifies the token that should be preserved, not the token that should be removed, contrary to other pattern-matching methods, such as de-id [11] that uses for instance known PHI, potential PHI and PHI indicator look-up tables. The approach of FASDIM is simpler, which enables to always value confidentiality over legibility of the de-identified text. As a consequence, FASDIM cannot be used as a pseudonymization tool, contrary to other methods.

FASDIM little relies on regular expressions, contrary to other pattern matching methods such as de-id [11]. We use only 48 regular expressions to remove titles, and then the experts are able to list simple number protection patterns. However, as FASDIM is mainly based on the removal of all the token that are not protected, and not the removal of specific patterns, more complex regular expression (such as address detection) are not necessary, which makes the software easier to maintain. In the evaluation study, no letter contains a complete address after de-identification, but 63.7% of the remaining PHIs are partial information about a place. Perhaps a more complete approach based on regular expression would enable to improve that point.

During the first step of FASDIM, the patients' names are optionally extracted from the HIS and removed from the text. This operation is a useful precaution, but is not sufficient, as there are frequently misspellings of names, and not indispensable, as the title patterns removal and the removal of unauthorized words may delete the names. The title patterns removal works well for formal discharge letters where titles (Mr, Mrs, Dr, etc.) are commonly used before person names, but might be less efficient for clinical notes that are less formal. On the other hand, when the family name is a common word (e.g. "Little"), the inappropriate disappearance of such a word may enable to guess the family name. This drawback can be sidestepped with pseudonymisation, which for instance consists in replacing the family names by other family names: indeed it can be decided not to replace family names that are also common names, but then the reader can't guess whether it is the original name or a pseudonym [34]. However, the names of the patients are more easily concealable in structured text.

As in most of the scientific papers dealing with de-identification evaluation, we have considered the patients' weights as "biometric identifiers" and therefore as PHIs (and 5.5% of our false negatives are patients' weights). This strict implementation may not be appropriate in medicine, as the patients' weights are important information, e.g. for obese or malnourished patients, or drug dose calculation. However, for a practical use of FASDIM, patients' weights could easily be conserved through the use of an appropriate number protection pattern as illustrated on Figure 4.

This work also introduces a new way to evaluate a de-identification method: the second evaluation phase estimates the proportion of medical information that has been preserved by the method. This point is important as the over-scrubbing of words has a variable importance depending on the word that is inappropriately removed. It demonstrates that the use of FASDIM would not alter the usability of the de-identified discharge letters for medical research or for activity-based payment systems.

Finally, as the FASDIM method is thought to be as language-independent as possible (cf. Objective 2), the same approach could probably be tested in other languages, although we cannot be sure it would be appropriate. However, such extension would be interesting as most of methods are designed only for English language.

VI. Conclusion

FASDIM is a fast and simple algorithm that enables to de-identify French free-text discharge letters. It preserves the patient confidentiality without threatening medical information. Is seems to be suitable especially when a medium corpus of letters has to be de-identified in a limited amount of time. Examples of source code and lists of words are freely available on the web [27]. The same method should be experimented and evaluated on other types of texts, including less formal texts (such as clinical notes). Its ability to work in other languages should also be evaluated.

References

[1] Summary of the HIPAA Privacy Rule http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary, last visited 27 November 2013.

[2] 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):246-52

[3] 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

[4] Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak. 2006;6:12

[5] Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research Arch. Pathol. Lab. Med. 2003;127(6):680-86

[6] Fielstein EM, Brown SH, Speroff T. Algorithmic De-identification of VA Medical Exam Text for HIPAA Privacy Compliance: Preliminary Findings; Medinfo. 2004

[7] Friedlin FJ, McDonald CJ. A Software Tool for Removing Patient Identifying Information from Clinical Documents. J Am Med Inform Assoc. 2008;15(5):601-10

[8] Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am. J. Clin. Pathol. 2004;121(2):176-86

[9] Grouin C, Rosier A, Dameron O, Zweigenbaum. Testing tactics to localize de-identification. Stud Health Technol Inform. 2009;150:735-39

[10] Morrison FP, et al. Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? J Am Med Inform Assoc. 2009;16(1):37-9

[11] Neamatullah I, Douglass MM, Lehman LH, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak. 2008;8:32

[12] Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. Medical document anonymization with a semantic lexicon. Proc AMIA Symp. 2000:729-33

[13] Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp. 1996:333-37

[14] Taira R, Bui A, Kangarloo H. Identification of patient name references within medical documents using semantic selectional restrictions. Proc AMIA Symp. 2002;757-61

[15] Thomas SM, et al. A successful technique for removing names in pathology reports using an augmented search an replace method. Proc AMIA Symp. 2002;777-81

[16] Velupillai S, Dalianis H, Hassel M, Nilsson GH. Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial. International Journal of Medical Informatics. 2009 Dec;78(12):e19-e26

[17] UMLS http://www.nlm.nih.gov/research/umls, last visited 27 November 2013.

[18] Aberdeen J, Bayer S, Yeniterzi R, et al. The MITRE Identification Scrubber Toolkit: Design, training, and assessment. International Journal of Medical Informatics 2010 Dec;79(12):849-59

[19] Aramaki E, Miyo K. Automatic Deidentification by Using Sentence Features and Label Consistency. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006

[20] Gardner J, Xiong L. HIDE: An Integrated System for Health Information De-identification. Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems 2008;254-9

[21] Hara K. Applying a SVM Based Chunker and a Text Classifier to the Deid Challenge i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006a

[22] Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc. 2007 Oct;14(5):574-580

[23] Uzuner Ö, Sibanda TC, Luo Y, Szolovits P. A De-identifier for Medical Discharge Summaries. Artif Intell Med. 2008 Jan;42(1):13-35

[24] Wellner B, Huyck M, Mardis S, et al. Rapidly Retargetable Approaches to De-identification in Medical Records. J Am Med Inform Assoc. 2007;14(5):564-73

[25] Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, Marsolo K, Jegga A, Kaiser M, Stoutenborough L, Solti I. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc. 2013 Jan 1;20(1):84-94. doi: 10.1136/amiajnl-2012-001012. Epub 2012 Aug 2.

[26] Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. BMC Med Res Methodol. 2012 Jul 27;12:109.

[27] The FASDIM web page: http://www.fasdim.org, last visited 27 November 2013

[28] Tu K, Klein-Geltink J, Mitiku TF, Mihai C, Martin J. De-identification of primary care electronic medical records free-text data in Ontario, Canada. BMC Med Inform Decis Mak. 2010;10:35

[29] Uzuner Ö, Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. J Am Med Inform Assoc. 2007;14(5):550-563

[30] Douglass MM, Clifford GD, Reisner A, Long WJ, Moody GB, Mark RG. Computer-Assisted De-Identification of Free-text Nursing Notes. 2005;32 :331-334

[31] Anatomical Therapeutic Chemical Classification System http://www.whocc.no/atc_ddd_index/, last visited 27 November 2013.

[32] International Classification of Diseases http://www.who.int/whosis/icd10, last visited 27 November 2013

[33] French common classification of medical procedures http://ccam.ameli.fr, last visited 27 November 2013

[34] Neubauer T, Heurix J. A methodology for the pseudonymization of medical data. Int J Med Inform. 2011 Mar;80(3):190-204.

[35] Suominen H, Lehtikunnas T, Back B, Karsten H, Salakoski T, Salanterä S. Applying language technology to nursing documents: pros and cons with a focus on ethics. Int J Med Inform. 2007 Oct;76 Suppl 2:S293-301.

[36] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc. 2007 Sep-Oct;14(5):550-63. Epub 2007 Jun 28.

[37] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000 Jun 13;101(23):E215-20.