Automated Classification of Free-text Pathology Reports for Registration of Incident Cases of Cancer

V. Jouhet[1]; G. Defossez[1]; A. Burgun[2]; P. le Beux[2]; P. Levillain[3,4]; P. Ingrand[1,5]; V. Claveau[6]
[1]Unité d'épidémiologie, biostatistique et registre des cancers de Poitou-Charentes, Faculté de médecine, Centre Hospitalier Universitaire de Poitiers, Université de Poitiers, Poitiers, France;
[2]INSERM U936, Faculté de médecine, Université de Rennes 1, Rennes, France;
[3]Anatomie et cytologie pathologiques, Centre Hospitalier Universitaire de Poitiers, Poitiers, France;
[4]Centre de Regroupement Informatique et Statistique en Anatomo-Pathologie de Poitou-Charentes, Faculté de médecine, Université de Poitiers, Poitiers, France:
[5]INSERM, CIC 802, Poitiers, France;
[6]IRISA – CNRS UMR 6074, Rennes, France

Keywords
Medical Informatics, neoplasm, pathology, free text, automated classification

Summary
Objective: Our study aimed to construct and evaluate functions called "classifiers", produced by supervised machine learning techniques, in order to categorize automatically pathology reports using solely their content.
Methods: Patients from the Poitou-Charentes Cancer Registry having at least one pathology report and a single non-metastatic invasive neoplasm were included. A descriptor weighting function accounting for the distribution of terms among targeted classes was developed and compared to classic methods based on inverse document frequencies. The classification was performed with support vector machine (SVM) and Naive Bayes classifiers. Two levels of granularity were tested for both the topographical and the morphological axes of the ICD-O3 code. The ability to correctly attribute a precise ICD-O3 code and the ability to attribute the broad category defined by the International Agency for Research on Cancer (IARC) for the multiple primary cancer registration rules were evaluated using F1-measures.
Results: 5121 pathology reports produced by 35 pathologists were selected. The best performance was achieved by our class-weighted descriptor, associated with a SVM classifier. Using this method, the pathology reports were properly classified in the IARC categories with F1-measures of 0.967 for both topography and morphology. The ICD-O3 code attribution had lower performance with a 0.715 F1-measure for topography and 0.854 for morphology.
Conclusion: These results suggest that free-text pathology reports could be useful as a data source for automated systems in order to identify and notify new cases of cancer. Future work is needed to evaluate the improvement in performance obtained from the use of natural language processing, including the case of multiple tumor description and possible incorporation of other medical documents such as surgical reports.

Correspondence to:
Vianney Jouhet
Unité d'épidémiologie, biostatistique et registre des cancers de Poitou-Charentes
Faculté de médecine
Centre Hospitalier Universitaire de Poitiers
Université de Poitiers
6, rue de la milétrie BP 199
86034 Poitiers Cedex
France
E-mail: vianney.jouhet@gmail.com

Methods Inf Med 2012; 51: 242–251
doi: 10.3414/ME11-01-0005
received: January 14, 2011
accepted: May 30, 2011
prepublished: July 26, 2011

1. Introduction

Most clinical databases in use today are special-purpose and restricted, and were not designed for interoperability [1]. In a White Paper by Safran et al., the secondary use of health data is discussed [2]. Secondary use of health data applies personal health information for uses outside direct health care delivery, which include, among other activities, research, quality and safety measurement, and public health. Despite the development of clinical repositories (e.g. [3]), secondary use of medical data for research and public health remains a challenge [4]. Kohane [5] contrasts the ease of access and sharing of biological data, which is mainly structured and standardized data, with the difficulty of utilizing patient information generated from routine care procedures, which is often in an unstructured free-text format. Integrating structured and unstructured information remains a challenge [6].

Pathology reports form one of the essential sources of data in order to include cases in cancer registries. These documents provide histological evidence of cancer on which the inclusion is mainly based. It is in many cases the pathologist who, via the analyses performed, confirms diagnosis and establishes the type of tumour, and when available, information contained in pathology reports is the most reliable source of data for the registration of incident cases of cancer [7].

For this registration procedure, a large part of the task involves classification targeting anatomical topography and tumour morphology. In order to harmonise data

collection, the International Association for Cancer Registries (IACR), the International Agency for Research on Cancer (IARC) and the World Health Organisation (WHO) specify that registered cases should be coded according to International Classification of Disease in Oncology, 3rd edition (ICD-O3) [8, 9]. Further to this, recommendations have been issued in collaboration with IACR, IARC, WHO and European Network of Cancer Registries (ENCR) concerning the registration rules for multiple primary cancers [10]. These recommendations specify the topographies and tumour morphologies that are to be recorded separately for one and the same individual so as to enable data comparisons across different populations. The recommendations define when a record should be considered to contribute to a new case or when it contributes to an already registered case, and the level at which data are to be agregated for follow-up of incidence and survival data.

As in numerous fields, the mass of information available in cancer registries is constantly increasing, so that manual processing is both tedious and costly. Automated cancer registration seems attractive as it may help to reduce delays in data production and allow personnel to devote more time to analysis and research. One of the major challenges is the development of automated analysis procedures for the data available so as to submit tumours for registration to the manual validation phase with the most plausible typology, defined both by anatomical site and histology.

Free-text pathology reports, often available in digital form, can be used for this data processing, but are not often available with an underlying terminology for anatomical sites and histology.

The Cancer Text Information Extraction System (caTIES) [11] has been developed in the framework of the caBIG project with focus on information extraction from pathology reports. Specifically, caTIES extracts information from free text Surgical Pathology Reports (SPRs), using the NCI Metathesaurus to populate caBIG-compliant data structures. Evaluation of CaTIES performance shows great precision [11] but recall has only been evaluated on the ability to retrieve ovarian cancer from radiology reports [12]. The Medical Text Analysis System/Pathology (MedTAS/P) [13] instantiates the Cancer Disease Knowledge Representation Model (CDKRM). MedTAS/P extracts cancer disease characteristics such as anatomical sites and histology from pathology reports. This system was evaluated on reports of colon cancer and showed good performance in terms of both recall and precision. However, to our knowledge, there is no efficient system enabling automatic extraction of tumour type for a tumour described in a French pathology report, and no study has evaluated recall and precision for more than one localisation.

Free-text pathology reports were also used by way of the stage coding in lung cancer [14, 15]. The CaFE (Case-Finding Engine) identifies cases of cancer in sets of pathology reports in English. This instrument, using lists of predefined terms, demonstrates a recall of 1 but a precision of 0.85 [16]. In the biomedical field, a large body of research has attempted to "map" biomedical concepts in clinical documents [17–19]. These approaches, mainly based on term-matching, have two main purposes: indexation, and the development of meta-data for use in automatised systems.

Little research has sought to apply automatic text classification (or categorisation) techniques on pathology reports to extract revelant informations. Li et al. [20] evaluated it over 203 pathology reports of colorectal cancer. To our knowledge, no study evaluated automatic text classification methods over more than one localisation. The automatic classification of text consists in annotation of the text so as to attribute a category solely on the basis of its content [21]. Since the 1990s, this discipline has developed considerably, and performances have improved, in particular since the development of machine learning techniques. One of the key resources for using learning machines is the availability of annotated documents [21]. These documents can be used in a learning process to construct a function by way of inference that enables categorisation of a new and unknown document. In this case the learning machine is said to be "supervised".

One of the tasks performed by physicians in the cancer registry in Poitou-Charentes (west France) consists in the annotation of pathology reports that contribute to the process of coding tumours. This provided us with the required database for the implementation of a supervised machine learning process for the classification of pathology reports in cancer units.

The aim of this study was to use pre-annotated data to construct and evaluate functions known as "classifiers", enabling the automatic categorisation of pathology reports in cancer unit solely from their textual content. Different classifiers were constructed, using supervised learning machines, according to two granularities, thus enabling two operational objectives to be envisaged:

- the attribution of a category defined by the 2004 recommendations on registration of multiple primary cancers [10]: generic anatomical site (IARC topography) and generic histology (IARC morphology)
- the attribution of a ICD-O3 code so as to provide a precise typology for the tumour: complete topography and complete morphology.

2. Methods

2.1 Classification Targets

The target class is defined as the annotation decided on by the human annotator for a given level of granularity. The ICD-O3 classification has two axes (▶Fig. 1), topography and morphology. For each, two levels of granularity were studied so as to comply with the two operational objectives of the annotation. Thus the granularities targeted were:

2.1.1 Topography

2.1.1.1 Coding of the complete topography according to ICD-O3

In ▶Figure 1, complete topography corresponds to C50.2 (upper inner quadrant of the breast)

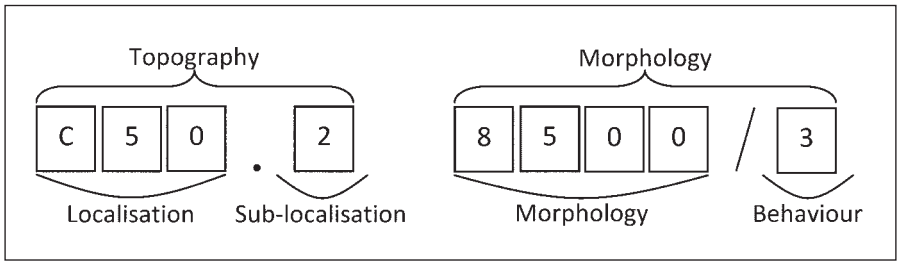

Fig. 1 Structure of the ICD-O3 code (with example of an Infiltrating duct carcinoma on the upper inner quadrant of the breast)

2.1.1.2 Coding according to the recommended level for multiple primary cancers (IARC topography)

This level comprises 54 target classes of the 330 classes in the complete topography. For certain tumour morphologies (Kaposi sarcoma and tumour of the haemato-poietic system), a single tumour is registered, independently from topography. In Figure 1, IARC topography corresponds to breast.

2.1.2 Morphology

2.1.2.1 Coding of the complete morphology according to ICD-O3

In ▶Figure 1, complete morphology corresponds to 8500/3 (Infiltrating duct carcinoma).

2.1.2.2 Coding according to the level recommended for multiple primary cancers (IARC morphology).

This level comprises 17 target classes of the 553 in the complete morphology. It corresponds to an adaptation of the morphology groups defined by Berg [22]. In ▶Figure 1, IARC morphology corresponds to adenocarcinoma.

2.2 Data Used

This study was conducted on data available in the Poitou-Charentes cancer registry for the year 2008. The collection and analysis of medical data by the cancer registry received the approval of the French regulatory authorities. The reports were written in French. Four main elements require prior definition:

- the distribution of topographies and tumour morphologies,
- the sources of the pathology reports,
- the manual production of annotations by physicians in the registry,
- and the relationship between the reports and ICD-O3 codes in the Poitou-Charentes cancer registry database.

The incidence of tumours according to localisation and tumour type is very uneven. For instance, in 2005, out of an estimated 320 000 cases in France, breast cancer, prostate cancer, colorectal cancer and lung cancer accounted for 37%, 34%, 12% and 10% of cases respectively, while pancreatic cancer, which ranks tenth out of 25, accounted for 2% of incident cases [23]. The data available for each targeted class relied on cases registered in the Poitou-Charentes cancer registry and the numbers of documents available was therefore varying a lot among the targeted class.

The pathology reports are in the form of free text in French, they are collected routinely from all the pathology laboratories in the Poitou-Charentes region, and annotated by the registry physicians. ▶Figure 2 gives an example of a report along with its translation. In the Poitou-Charentes registry database, one individual can have several tumours, and one tumour can be described in several reports. Likewise, it should be noted that one and the same report may contain the descriptions of several tumours. In addition, the reports describing secondary tumours (lymph glands or bowel) are directly linked to the primary tumour.

Finally a considerable proportion of the reports are made up of complementary notes to be attached to the earlier report (immuno-histo-chemical analyses and other expertise). These reports are linked to the tumours that they describe, but are not attached to the reports to which they contribute.

In this setting, and so as to ensure a single link between the report and its annotations, we included all pathology reports of individuals meeting the all following criteria:

1. a single identified tumour
2. an invasive tumour (behaviour coded /3 in ICD-O3)
3. a tumour having a manually validated, complete ICD-O3 coding
4. a tumour described by at least one pathology report
5. no identified organ or lymph node metastatic site

All the steps to prepare and represent the reports were performed using SAS v9.1. The construction and the evaluation of the different "classifiers" was performed on WEKA v3.6.2, a program developed in Java which provides classification tools [24].

2.3 Preparation of the Reports

Pre-processing of the reports had to take into account their heterogeneous forms, depending on the laboratory issuing them and even the pathologist who drafted the report. All the reports available for a given tumour were concatenated. Non-useful characters were replaced by spaces (special characters, parentheses and brackets, figures and operators). Non-informative words (e.g. certain articles and pronouns such as le, la, lui, elle) were removed using a "stop list". Lower case was used for all words in all reports, and accents, variably used, were removed. Finally a stem identification process removed plural and feminine forms [25].

2.4 Construction of the Training and Test Sets

To enable the evaluation of the automatic classification, the available data were split into two separate sets: a training set and a test set. Only the training set is used in all the construction phases of the classifier [21]. Separation into a training set and a test set was performed independently for

each level of granularity. Depending on the prevalence of reports available for each target class in the granularity being processed, the reports were allocated to one or other of the sets of data. The reports were randomly allocated, 75% to the training set and 25% to the test set. We set the minimum at 25 reports for the training set and 5 for the test set, i.e. at least 30 reports were available for a given class. The reports belonging to target classes comprising smaller numbers than this were re-allocated to a class labelled "Others". With regard to IARC topography, morphologies for which the topography was non-informative (Kaposi sarcoma and tumours of the haematopoietic system) were grouped in a class labelled "systemic tumour".

2.5 Representation of Reports

In order to perform automatic classification tasks, a representation of the reports has to be produced, in a form that can be interpreted by the learning algorithms. The most widely used representation is the projection of the report descriptors (basic forms that will be used for the purpose of representation, for instance words) within a vector space [26, 27]. Here the documents are represented by a vector, for which the number of dimensions corresponds to the number of different descriptors identified from the set of documents.

The representation of the reports requires three successive procedures [21]:

- The choice of descriptors
- The reduction of the number of dimensions (optional)
- The choice of a quantifiable representation.

RENSEIGNEMENTS CLINIQUES :
Masse de plus de 2 cm de découverte récente, à l'union des QS du sein droit.
ACR5.
2 prélèvements 16 G.
MICROSCOPIE : MICROBIOPSIES MAMMAIRES DROITES (UNION DES QS) (2 carottes de 7 et 9 mm)
Les 2 carottes biopsiques examinées sont lésionnelles et d'aspect microscopique superposable.
Il existe en effet une prolifération cellulaire agencée en travées dissociant largement les tissus fibro-collagéniques présents.
Les cellules qui constituent cette prolifération montrent une anisocaryose, quelques mitoses.
Les immunomarquages pratiqués montrent :
- P63 négative en périphérie des structures épithéliales proliférantes, versus témoin interne positif.
- Ecadhérine légérement positive.
Ces aspects morphologiques correspondent à un CARCINOME CANALAIRE INFILTRANT dont le SBR est estimé à 2 dans la limite du matériel lésionnel examiné.
A noter la présence d'un embole vasculaire péri-tumoral.

CLINICAL INFORMATION
Recently discovered mass of over 2cm in diameter at junction between the upper quadrants of the right breast.
ACR5.
Two samples taken 16G.
MICROSCOPIC EXAMINATION: RIGHT BREAST MICRO-BIOPSIES (JUNCTION UPPER QUADRANTS) (2 core samples of 7 and 9 mm)
The two biopsy core samples examined are lesional and microscopic aspect is identical.
There is a cellular proliferation in bands cutting across the fibro-collagen tissues present.
The cells composing this proliferation demonstrate anysocaryosis, and some mitoses.
The immuno-marking performed showed:
Negative P63 on the periphery of the proliferating epithelial structures, versus positive internal control.
Slightly positive Ecadherine
These morphological aspects correspond to an INFILTRATING DUCT CARCINOMA, for which SBR is estimated to be 2 on the basis of the lesional material examined.
The presence of a peri-tumoral vascular embolus can also be noted.

Fig. 2 Example of a French pathology report and its English translation for the purpose of the article

2.5.1 Choice of descriptors and dimensionality reduction

We chose a representation taking the form of a "bag of words". This is the most classic and simplest representation, whereby the text is divided up into the words that it comprises [21]. Some of these words will be used as dimensions for the vector representing the document. The number of dimensions (words) extracted from the training data is potentially very large (easily 10,000, even for texts of only moderate length). To reduce processing time, and also to avoid "overfitting" [21, 28], the number of dimensions was reduced by term selection using the multivariate Chi^2 method [29]. This method selects the terms that are the most strongly related to each of the target classes.

2.5.2 Quantifiable representation of descriptors

The aim here is to represent the report by a vector of numerical descriptors. We tested different methods of representation:

- Classic methods
  - *tf* for "Term Frequency", where it is considered that the more often a term occurs in a document, the more representative it is of text content, so

$$w(t_k, d_j) = \#(t_k, d_j).\log\frac{|D|}{\#T_r(t_k)}$$

Equation 1 *tf.idf*

that its weight should be of a high magnitude [21]

– *tf.idf* for Term Frequency with Inverse Document Frequency, where it is considered that a term that appears in numerous documents in the training corpus has little discriminating power for the task of classification, and should have a correspondingly small weight [21, 27, 29, 30]. The weight of a descriptor t_k in a document d_j, noted $w(t_k,d_j)$, is given by ▶Equation 1 [30] where |D| is the total number of documents, $\#(t_k,d_j)$ is the number of occurrences of term t_k in document d_j and #Tr(t_k) is the number of documents in which t_k appears (see ▶Eq. 1). Given the imbalance among classes in the training data available, we made the hypothesis that the use of *tf.idf* was liable to decrease the weight of the terms associated with classes in which number of training documents available were large. We therefore developed and evaluated a new method of representation.

- *tf.icf* for Term Frequency with Inverse Class Frequency. This method proposes to use the relationship between the terms and the classes in the corpus in order to generate a weighting index. This index, which we call "*icf*" depends on the distribution of a term across the different classes in the corpus. It is considered that a term with a homogenous distribution across classes is not discriminant for the task of classification, and should therefore have a correspondingly small weight. ▶Equation 2 shows an application of the *tf.icf* function, where $w(t_k,d_j)$ is the weight obtained by applying the *tf.icf* function, |C| is the total number of classes, $\#T(C_i)$ is the number of term in the class C_i, $\#(t_k,C_i)$ is the number of occurrences of the term tk in class C_i.

For the representation of the descriptors to range from 0 to 1, the weight of each descriptor is standardized by l^2-norm of the document vector (▶Eq. 3) [21] where $w(t_k,d_j)$ is the weight of a descriptor k in a document j calculated using tf.idf or *tf.icf* and $|T|$ is the total number of terms in the document.

2.6 Classification

The classification was performed using machine learning techniques. The learning is said to be "supervised", since both the reports and their corresponding coding were available for all the data used in the development and validation stages. We compared two learning algorithms commonly used in classification tasks applied to text:

- the Naïve Bayes classifier [21, 31]
- the Support Vector Machine (SVM) classifier [21, 32, 33].

2.7 Evaluation

The classification was evaluated on the test data set. The documents in the test sets were represented and classified according to the parameters calculated and the classifier constructed from the training data. The results of this classification were compared with the manual annotations of the registry physicians. Three indicators were used:

- Percentage of correctly classified reports
- Mean F1-measure (harmonic mean of precision and recall [21])
- Kappa agreement coefficient [34]

These measures are classically used to assess the performance of classifiers, in the field of automatic classification (F1-measure), or for agreement between two methods, in particular in the biomedical field (Kappa coefficient).

All these steps were reiterated 10 times. Each time a training set and a test set were generated randomly from the overall report corpus.

A qualitative assessment of error was performed by way of systematic perusal of the wrongly classified reports. The main sources of error are discussed in Section 3.2.

$$\text{weight}(t_k, C_r) = \frac{\frac{\#(t_k, C_r)}{\#_T(C_r)}}{\max_{i=1}^{|C|}\left(\frac{\#(t_k, C_i)}{\sum_{j=1}^{\#_T(C_i)}\#(t_j, C_i)}\right)}$$

$$w(t_k, d_j) = \#(t_k, d_j).\log\frac{|C|}{\sum_{i=1}^{|C|}\text{weight}(t_k, C_i)}$$

Equation 2 *tf.icf*

3. Results

▶Tables 1 and 2 present the detailed results for each representation, classification method and granularity used.

$$r_{kj} = \frac{w(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|}(w(t_s, d_j))^2}}$$

Equation 3 Normalization (r_{kj}) of a descriptor k in a document j

In all, the complete corpus of data comprised 5121 documents. The reports used were derived from 16 laboratories and 35 pathologists.

After stemming and "stop words" suppression, 10193 unique terms were identified in the overall corpus. This corresponds to the maximum dimensionality of representation's vectors before applying the dimensionality reduction (e.g. the maximum final dimensionality after reduction for the IARC morphology would be 8 targeted classes × 100 = 800 terms).

The comparison of performance of the different methods of representation and classification is provided in the annexes. In the experimentations performed overall, SVM was the classifier that demonstrated the best performances. Representations of the tf type provided the least satisfactory results. The two other methods of representation were comparable, with slightly better results for tf.icf (▶Table 3).

We chose to use the most efficient method across all the granularities targeted. The stages in this model were: a representation of the descriptors of the tf.icf type and a classifier of the SVM type. The following results present the quantitative and qualitative evaluation of this method for the different granularities targeted.

Table 1 Classification among granularities of the tumoral topography

Representation and model	Correct (%)	F1-measure	Kappa
IARC topographic class (16 categories)			
TF			
Naïve Bayes	85.5	0.802	0.799
SVM	96.4	0.963	0.952
TF.IDF			
Naïve Bayes	88.4	0.844	0.843
SVM	96.6	0.966	0.955
TF.ICF			
Naïve Bayes	88.8	0.848	0.847
SVM	**96.7**	**0.967**	**0.957**
Complete topography (26 categories)			
TF			
Naïve Bayes	60.9	0.507	0.486
SVM	72.4	0.713	0.655
TF.IDF			
Naïve Bayes	65.5	0.567	0.556
SVM	**72.6**	**0.716**	**0.659**
TF.ICF			
Naïve Bayes	65.7	0.569	0.559
SVM	72.5	0.715	0.657

3.1 Quantitative Evaluation

Overall, good consistency was observed among the different evaluation measures implemented. However, regarding IARC morphology, there is a discrepancy between the F1-measure (demonstrating the best values for the granularities overall) and the slightly lower kappa values.

Irrespective of the axis considered (topography or morphology) the classification at the level recommended for multiple primary cancer coding showed good performances, with F1-measures at 0.967 and kappa coefficients of 0.957 to 0.892 for topography and morphology respectively. In contrast, the classification according to the ICD-O3 level of precision proved more delicate, in particular for the topography axis (▶Table 4).

▶Tables 5 and 6 give details of the quantitative indices for each target class in IARC topography and Berg's morphological groups.

Unsurprisingly, the "Others" class systematically showed a low F1-measure, while the classes with the largest prevalence available for the training process showed better performances. A tendency was noted towards a decrease in the F1-measure proportionately to the decrease in prevalence available for training. However, precision was systematically greater than 0.7 for all the target classes studied (with the exception of the "Others" for the two axies and "Unspecified type of haematopoietic and lymphoid tissue" classes in IARC morphology); the small prevalence in the training set tended above all to alter recall. The examination of confusion matrices showed that the most frequently encountered errors consisted in allocation of a report belonging to a class with small prevalence to a class comprising large prevalence.

Regarding IARC topography, "Colon" and "Rectum and rectal-sigmoid junction" showed poorer performances independently of the prevalence of training data available. They correspond to difficulties to distinctly identify these two contiguous topographies of the lower digestive tract, as pointed out during the qualitative examination of errors.

3.2 Qualitative Evaluation of Classification Errors

3.2.1 Reports Describing Several Examinations

Certain situations can occur in which a single report concerns the examination of several samples. Most often, this relates to series of samples taken within a single diagnostic phase, examined at the same time and included in a single pathologist's report. For instance, for the diagnosis of

Representation and model	Correct (%)	F1-measure	Kappa
IARC morphological class (8 categories)			
TF			
Naïve Bayes	91.8	0.893	0.639
SVM	97.0	0.965	0.887
TF.IDF			
Naïve Bayes	94.0	0.924	0.759
SVM	97.0	0.966	0.887
TF.ICF			
Naïve Bayes	94.5	0.929	0.782
SVM	**97.1**	**0.967**	**0.892**
Complete morphology (18 categories)			
TF			
Naïve Bayes	75.8	0.683	0.646
SVM	86.3	0.853	0.812
TF.IDF			
Naïve Bayes	77.9	0.710	0.685
SVM	86.3	0.852	0.811
TF.ICF			
Naïve Bayes	78.1	0.714	0.691
SVM	**86.4**	**0.854**	**0.813**

Table 2 Classification among granularities of the tumoral morphology

Table 3 Comparison of performance by tf.idf and tf.icf according to granularity using a SVM-type classifier

Granularities	F1-measure mean (sd)*		Kappa	
	tf.idf	tf.icf	tf.idf	tf.icf
Topography				
IARC**	0.965 (0.003)	0.967 (0.003)	0.956	0.957
Complete topography	0.716 (0.008)	0.715 (0.008)	0.659	0.657
Morphology				
IARC**	0.966 (0.003)	0.967 (0.003)	0.887	0.892
Complete morphology	0.852 (0.007)	0.854 (0.006)	0.811	0.813

* Mean and standard deviation over the ten iterations
** Coding according to level recommended for registration of multiple primary cancers

cancers of the digestive tract, a colonoscopy gives rise to several samples being taken and can often be accompanied by upper gastrointestinal endoscopy which will also generate sampling procedures. In these reports, descriptions of healthy and cancerous tissue can coexist. For each of these descriptions there is a corresponding topography, and the human annotator in this case retains the topography relating to the tumour. The automatic classification of the topography may however retain a topography that does not correspond to the tumour.

3.2.2 Semantic Influence of a Corpus-based Training Data

As our method uses corpus-based training data, the classifier learned the strong association that some specific terms are classically used by majority of pathologists to describe cancer of certain anatomical location. As an influence in testing, an unspecific term that deviates from the classical usage may cause classification error, even if the unspecific term implies the associated location. This is the case for instance for the term "spinocellular epithelioma", which is classically used to describe epidermoid carcinoma of the skin. On account of this usage, there was a very strong association in the training data between this term and a cutaneous localisation. In the test set, the use by a pathologist of the term "epidermoid carcinoma" for the skin could thus lead to classification errors.

4. Discussion

The system of classification according to the level recommended by IARC for the registration of multiple primary cancers demonstrated very good performance. However the complete encoding of the tumour at a level of detail as refined as that of the ICD-O3 remains a delicate task. This is particularly true for complete topography, but we also need to take into account the fact that the pathologist does not always possess the information concerning the exact topography of the tumour, so that this data is not always present in the reports, and the practitioner who codes the tumour may, if needed, have used information derived from another document.

Our method of representation (*tf.icf*) showed the best performance. It introduces the notion of relationships between terms and classes for quantitative representation, and reduces the effects of the imbalance between classes in the training process. Although differences observed were small, an improvement in performance was nonetheless noted with icf, and this method should therefore be evaluated in different settings and on different types of textual data. Lertnattee et al. investigated inverse

Table 4 Performance of the classification according to granularity

Granularity	Number of classes	Correct (%)	F1-measure	Kappa
Topography				
IARC*	16	96.7	0.967	0.957
Complete topography	26	72.5	0.715	0.657
Morphology				
IARC*	8	97.1	0.967	0.892
Complete morphology	18	86.4	0.854	0.813

* Coding according to level recommended for registration of multiple primary cancers

class frequency in centroid-based text classification [35]. The functions applied were based on the number of classes that contained the term to weight and neither used the number of occurrences of this term nor the total number of terms in each targeted class. By introducing these two factors, our representation method tried to capture the distribution of terms among classes taking into account the size of the training set of each class.

Our evaluation required the production of reliable indicators in view of our objective. The selected reports formed a corpus of simple cases for which no secondary localisation or multiple primary sites had been identified. This made it possible to ensure that the reports contained conclusions for only one topography and one invasive morphology (according to the 2004 recommendations). This process is obviously unrealistic from the point of view of the routine functioning of registries. We left out 16% of reports with multiple primary localizations or organ or lymph node metastatic site identified. In addition, for the topographic axis, we noted the potential impact of the coexistence of tumoral and non-tumoral pathologies on report content. It is therefore essential to identify reports containing several topographies. These elements strongly suggest the potential benefit of a multi-label classification, so that several target classes could be allocated to a single report. Both Naïve Bayes and SVM can generate ranked class predictions and can be adapted for multi-label classification.

The reports were derived from a large sample of pathologists working in different laboratories. The dataset therefore reflected certain variability in the drafting. This suggests that the introduction of written reports from other pathologists should not generate extra difficulties, so long as reports from these pathologists are available for the training process.

The use of different quantitative indices enabled apprehension of the overall performances of the different classifiers. The coherence of the results in relation to IARC topography, complete topography and complete morphology confirmed a certain level of performance. However the examination of disagreement concerning IARC morphology puts some perspective on the high F1-measure. The examination of the confusion matrices concerning this granularity showed that most of the mistakes concerned false positives of the "adenocarcinoma" class. These false positives in fact amounted to only a small prevalence in re-

Table 5 Detailed results for topography at level recommended by IARC for registration of multiple primary cancers

IARC topographic class	Training	Test	Recall	Precision	F1-measure
Prostate	1529	509	0.999	0.998	0.999
Breast	972	324	1.000	0.995	0.998
Systemic tumour	264	88	0.990	0.976	0.983
Skin	144	48	0.969	0.973	0.970
Uterus	54	18	0.917	0.928	0.922
Colon	405	134	0.940	0.902	0.921
Liver and intrahepatic bile duct	44	14	0.950	0.888	0.917
Ovary	26	5	0.900	0.930	0.910
Trachea, bronchi, lung	39	12	0.883	0.932	0.905
Oesophagus	44	14	0.850	0.934	0.887
Gallbladder and other parts of biliary duct	30	9	0.789	0.928	0.844
Pancreas	25	5	0.880	0.812	0.831
Rectum and rectal-sigmoid junction	166	55	0.802	0.850	0.823
Stomach	31	10	0.730	0.919	0.806
Female genital organs	26	5	0.660	0.883	0.728
Other	54	18	0.639	0.692	0.661

Table 6 Details of results for morphology at level recommended by IARC for registration of multiple primary cancers

IARC morphological class	Training	Test	Recall	Precision	F1-measure
Adenocarcinomas	3240	1079	0.997	0.984	0.990
Squamous and transitional cell carcinomas	244	81	0.944	0.967	0.956
B-cell neoplasms	179	59	0.958	0.852	0.902
Hodgkin lymphomas	25	8	0.850	0.947	0.889
Other specific carcinomas	63	21	0.743	0.852	0.791
T-cell and NK-cell neoplasms	25	8	0.500	0.880	0.627
Other	42	14	0.357	0.693	0.461
Unspecified types of haematopoietic and lymphoid tissues	25	8	0.113	0.430	0.167

lation to the true positives in the adenocarcinoma class. Thus the effect on precision and on the F1-measure was very small. Adenocarcinomas amounted to nearly 87% of Berg's morphologies. As a result of weighting procedures, this class had a preponderant role in the calculation of the global F1-measure. The kappa coefficient, in contrast, may underestimate concordance in this situation. As Kappa is a chance-corrected measure [36], it is affected by a skewed prevalence. In this situation, the probability of chance agreement on high prevalence cases is very high [36], this is the "Kappa paradox" [37, 38] and a low kappa measure may not necessarily reflect a low concordance [39]. Its fairly high value here reflected adequate performances, although they were below what the F1-measure might lead one to expect.

Difficulties of a semantic nature were identified. They highlight the limitations of using words as descriptors, and point to the need for processing of the classification upstream. The use of UMLS (Unified Medical Language System) and "mapping" of reports by way of tools such as "MetaMap" [17] in order to extract and use medical concepts as descriptors of the reports should enable the impact of these problems to be reduced. This approach would also enable identification of concepts described by several successive words (e.g."infiltrating duct carcinoma"). These words are at present treated as independent entities. In addition, concept-based feature aggregation could reduce feature dimensionality and remove dependent features simultaneously. However as yet no equivalent tool has been developed in French.

Moreover, text information extraction systems based on the NCI Enterprise Vocabulary System or the UMLS make use of the sets of synonyms and the semantic categorisation provided by these resources. In caTIES, concepts are first identified using MetaMap. Then after detecting those that are explicitly negated, concepts are categorized on the basis of vocabulary semantic types, and classified as Diagnosis, Procedure or Organ type. Evaluations show a high level of precision (0.94 over 30 different queries ranging from low to high complexity). This suggests that, if equivalent resources were available for French, these methods could be adapted to extract and qualify concepts before applying our categorisation methods using a "bag of concepts" representation.

The performance metrics using CaTIES show a 0.98 precision in retrieving cases of prostatic adenocarcimoma identified on a prostatectomy for 60–80-year-old men [11]. Our results are comparable (0.998 for prostate and 0.984 for adenocarcinomas). However, the recall was not evaluated in the study by Crowley et al. [11] and this measure remains essential in a registry setting to ensure exhaustiveness. CaTIES was used to code free-text radiology reports in order to retrieve reports describing ovarian cancers with a 0.82 recall [12]. Our method provided a 0.90 recall on pathology reports for ovary topographic class (table V). Li et al. proposed a method to extract information from pathology reports using machine learning algorithms. They achieved the best performances with feature selection and Naive Bayes with a 0.511 F1-measure for tumour site and 0.964 for microscopic type [20].

The prevalence of reports available for the training process was very variable according to the target classes, on account of the natural variability of distribution of cancer localisations and tumour morphologies. The data available in our study did not enable all the classes for each granularity to be envisaged. It is however clear from our results that certain target classes are easier to annotate than others. Therefore this research work needs to be widened to datasets in which prevalence in the different classes could enable use of a larger number of target classes for each granularity. We noted the influence of this prevalence on the performance of classifiers. The threshold of 25 reports used for the training process was probably the reason for the decrease in performances, and it is probable that an increase in this threshold would enable an improvement.

Numerous other elements contribute to the coding of tumours. It is therefore clear that the sole use of pathology reports for automatic coding deprives users of part of the information that is available to a human annotator. Other text documents are often available, such as surgical reports, letters, or notes from pluri-disciplinary team meetings, and these could possibly be used in the same manner. For instance, surgical reports would provide a better source for the coding of complete topography.

The present results make it possible to envisage the incorporation of pathology reports as sources of data for the automated processing of information upstream of the manual validation phases in cancer registries. Adjustments to allow for semantic factors and for reports that describe several

samples should lead to the improvement of performance, and widen the field of application.

Acknowledgments

The authors would like to thank Pathologists of CRISAP Poitou-Charentes: Elisabeth Baltus , Dominique Battandier, François Baylac, Christine Blanchard, Françoise Bonneau-Hervé, Elisabeth Boudaud, Karine Boye, Philippe Debiais, Rony El Khoury, Catherine Emile, Catherine Fleury, Gaëlle Fromont Hankard, Jean-Michel Goujon, Harold Ip Kan Fong, Sylvain Labbé, Christian Lancret, Pierre Levillain, Didier Lhomme, Marie-Josée Loyer-Lecestre, Baudoin Mazet, Françoise Memeteau, Serge Milin, Olivier Nohra, Martine Paoli-Labbé, Dominique Petrot, Olivier Renaud, Mercédès Riols, Denis Roblet, Mokrane Yacoub who provided electronic full-text annotated pathology reports. They would also like to thank the medical records assistants, Nicolas Mériau and Sébastien Orazio of the *Registre des cancers de Poitou-Charentes* for their helpful contribution to the implementation of this research.

References

1. Maojo V, Kulikowski CA. Bioinformatics and medical informatics: collaborations on the road to genomic medicine? J Am Med Inform Assoc 2003; 10 (6): 515–522.
2. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007; 14 (1): 1–9.
3. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc 2007. pp 548–552.
4. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med 2009; 48 (1): 38–44.
5. Kohane IS. Bioinformatics and clinical informatics: the imperative to collaborate. J Am Med Inform Assoc 2000; 7 (5): 512–516.
6. Garcia-Remesal M, Maojo V, Billhardt H, Crespo J. Integration of relational and textual biomedical sources. A pilot experiment using a semi-automated method for logical schema acquisition. Methods Inf Med 2010; 49 (4): 337–348.
7. MacLennan R. Cancer registration: principles and methods. Items of patient information which may be collected by registries. IARC Sci Publ 1991; 1: 43–63.
8. Buemi A. Pathology of Tumours for Cancer Registry Personnel. IARC, Lyon; 2008.
9. Percy C, Fritz A, Jack A, Shanmugarathan S, Sobin L, Parkin D, et al. International Classification of Diseases for Oncology (ICD-O). 3rd ed. World Health Organization; 2000.
10. Curado M, Okamoto N, Ries L, Sriplung H, Young J, Carli M, et al. International rules for multiple primary cancers (ICD-O). 3rd ed. 2004.
11. Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. J Am Med Inform Assoc 2010; 17 (3): 253–264.
12. Carrell D, Miglioretti D, Smith-Bindman R. Coding free text radiology reports using the Cancer Text Information Extraction System (caTIES). AMIA Annu Symp Proc 2007. p 889.
13. Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform 2009; 42 (5): 937–949.
14. McCowan IA, Moore D, Fry MJ. Classification of cancer stage from free-text histology reports. Conf Proc IEEE Eng Med Biol Soc 2006; 1: 5153–5156.
15. McCowan IA, Moore DC, Nguyen AN, Bowman RV, Clarke BE, Duhig EE, et al. Collection of cancer stage data by classifying free-text medical reports. J Am Med Inform Assoc 2007; 14 (6): 736–745.
16. Hanauer D, Miela G, Chinnaiyan A, Chang A, Blayney D. The Registry Case Finding Engine: An Automated Tool to Identify Cancer Cases from Unstructured, Free-Text Pathology Reports and Clinical Notes. Journal of the American College of Surgeons. 2007; 205 (5): 690–697.
17. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001. pp 17–21.
18. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Stud Health Technol Inform 2004; 107 (Pt 1): 268–272.
19. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004; 11 (5): 392–402.
20. Li Y, Martinez D. Information extraction of multiple categories from pathology reports. Australasian Language Technology Association Workshop (ALTA Workshop 2010): Australasian Language Technology Association (Melbourne); 2010. pp 41–48.
21. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv 2002; 34 (2): 1–47.
22. Berg JW. Morphologic classification of human cancer. In: Shottenfeld D FJ, Jr., editor. Cancer epidemiology and prevention. 2nd ed. New York: Oxford University Press; 1996.
23. Belot A, Grosclaude P, Bossard N, Jougla E, Benhamou E, Delafosse P, et al. Cancer incidence and mortality in France over the period 1980–2005. Rev Epidemiol Santé Publique 2008; 56 (3): 159–175.
24. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl 2009; 11 (1): 10–18.
25. Savoy J. A stemming procedure and stopword list for general French corpora. J Am Soc Inf Sci 1999; 50 (10): 944–952.
26. Salton G, Wong A, Yang C. A vector space model for information retrieval. Communications of the ACM 1975; 18 (11): 613–620.
27. Laroum S, Béchet N, Hamza H, Roche M. Classification automatique de documents bruites a faible contenu textuel. RNTI: Revue des Nouvelles Technologies de l'Information 2009; 1: 25.
28. Yang Y, Jan P. A comparative study on feature selection in text categorization. In: Proceedings of ICML-97, 14th international conference on machine learning, Nashville, TN; 1997. pp 412–420.
29. Clech J, Rakotomalala R, Jalam R. Séléction multivariée de termes. XXXVèmes Journées de Statistiques. Lyon, France; 2003. pp 933–936.
30. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manage 1988; 24 (5): 513–523.
31. John GH, Langley P. Estimating Continuous Distributions in Bayesian Classifiers. Eleventh Conference on Uncertainty in Artificial Intelligence. San Mateo: Morgan Kaufmann; 1995. pp 338–345.
32. Boser BE, Guyon I, Vapnik NV. A training algorithm for optimal margin classifiers. 1992. pp 144–152.
33. Platt JC. Fast training of support vector machines using sequential minimal optimization. 1999. pp 185–208.
34. Fleiss JL. Statistical methods for rates and proportions. New York: Wiley; 1981.
35. Lertnattee V, Theeramunkong T. Analysis of inverse class frequency in centroid-based text classification. IEEE International Symposium on Communications and Information Technology 2004; 2: 1171–1176.
36. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005; 12 (3): 296–298.
37. Feinstein A, Cicchetti D. High agreement but low kappa: I The problems of two paradoxes. J Clin Epidemiol 1990; 43 (6): 543–549.
38. Cicchetti D, Feinstein A. High agreement but low kappa: II Resolving the paradoxes. J Clin Epidemiol. 1990; 43 (6): 551–558.
39. Viera A, Garret J. Understanding interobserver agreement: the kappa statistic. Fam Med 2005; 37 (6): 543–549.