Automated Selection of Relevant Information for Notification of Incident Cancer Cases within a Multisource Cancer Registry

V. Jouhet1,2; G. Defossez1; CRISAP3; CoRIM4; P. Ingrand1,5

1Registre général des cancers de Poitou-Charentes, Faculté de médecine, Centre Hospitalier Universitaire de Poitiers, Université de Poitiers, Poitiers, France; 2Centre de Regroupement Informatique et Statistique en Anatomie Pathologique, Faculté de Médecine, Université de Poitiers, Poitiers, France; 3Collège Régional de l’Information Médicale de Poitou-Charentes, Centre Hospitalier Général Georges Renon, Niort, France; 4INSERM, CIC 802, Poitiers, France

Keywords
Multisource information system, cancer registry, tumour notification, selection algorithm, relevant information

Summary
Objective: The aim of this study was to develop and evaluate a selection algorithm of relevant records for the notification of incident cases of cancer on the basis of the individual data available in a multi-source information system.

Methods: This work was conducted on data for the year 2008 in the general cancer registry of Poitou-Charentes region (France). The selection algorithm hierarchizes information according to its level of relevance for tumoral topography and tumoral morphology independently. The selected data are combined to form composite records. These records are then grouped in respect with the notification rules of the International Agency for Research on Cancer for multiple primary cancers. The evaluation, based on recall, precision and F-measure confronted cases validated manually by the registry’s physicians with tumours notified with and without records selection.

Results: The analysis involved 12,346 tumours validated among 11,971 individuals. The data used were hospital discharge data (104,474 records), pathology data (21,851 records), healthcare insurance data (7508 records) and cancer care centre’s data (686 records). The selection algorithm permitted performances improvement for notification of tumour topography (F-measure 0.926 with vs. 0.857 without selection) and tumour morphology (F-measure 0.805 with vs. 0.750 without selection).

Conclusion: These results show that selection of information according to its origin is efficient in reducing noise generated by imprecise coding. Further research is needed for solving the semantic problems relating to the integration of heterogeneous data and the use of non-structured information.

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

doi: 10.3414/ME12-01-0101
received: October 25, 2012
accepted: March 27, 2013
prepublished: April 24, 2013

1. Introduction

Cancer registries have the task of exhaustively recording incident cases of cancer in a given territory. In order to harmonise data collection, the International Agency for Cancer Registries (IACR), the International Agency for Research on Cancer (IARC) and the World Health Organization (WHO) specify that registered cases should be coded according to the International Classification of Diseases in Oncology, 3rd edition (ICD-O-3) [1, 2]. Further to this, recommendations have been issued in collaboration with IACR, IARC, WHO and the European Network of Cancer Registries (ENCR) concerning registration rules for multiple primary cancers [3]. These recommendations define when a record should be considered to contribute to a new case and when it should be considered to contribute to an already registered case, and the level at which data are to be aggregated for follow-up of incidence and survival data.

The job of cancer registries extends well beyond the mere recording of incident cases. In France, the National Committee of Registers (NCR) evaluation grids to be applied for the accreditation of registries include not only the methods used and the quality of the records, but also the use made of the data, the interest and the originality of the research work conducted [4].

To enable the registries to make full use of their expertise and research function in
the area of cancer epidemiology, the optimization of registration procedures for incident cases of cancer is crucial, and has been recalled in the two French national cancer plans. This optimization is in particular necessary for registries covering large territories and populations. The implementation of automated and semi-automated procedures to assist in detecting and documenting incident cases of cancer is therefore an attractive approach in this setting.

In order to be able to register cases, cancer registries staff, must be informed that possible new cases should be registered. This process, called notification, was historically based on voluntary practitioners that declared all new cases they encountered. As early as 1998, an IARC technical report was drawn up describing the methods used by different registries for the establishment of automated notification procedures [5]. To ensure adequate cover of a population of several million people, at a very early stage the Ontario registry was obliged to develop methods for data acquisition [6]. Both notification and record tasks in this registry are automated with no clerical interventions in routine processing [5, 7]. The integration of multi-source data into information systems that are structured around the patient makes it possible to optimize automatic processing for the notification and registration of incident cases, and the recording of complementary data [8]. The use of this accumulated information for case notification is a logical strategy in a perspective of exhaustiveness. This view has led registries to increasingly diversify notification sources. The mean number of notification sources has thus become a criterion for exhaustiveness, and a percentage of histological confirmation of cases that is too high should lead to suspicion of non-exhaustiveness [9]. The price to pay for this approach is excess notification of false incident cases following coding inaccuracies by the different data sources. These false cases require manual processing to be removed. Reducing their numbers without affecting exhaustiveness would enable time to be saved (and hence cost) for each case finally registered and would improve the quality of the automated data produced.

Olive et al. have presented a critical analysis of French hospital discharge data for the epidemiology of cancers. In particular, they noted difficulties relating to the use of data in isolation to detect incident cases [10]. This finding underlines the importance of using diverse sources for the notification of new incident cases. Deterministic [8, 11, 12] or probabilistic algorithms using artificial learning techniques [13] have been developed and evaluated. These works were mainly based on interrelating and aggregating data from different sources, but without any selection of relevant information according to its source. Couris et al. used coded information on surgical procedures performed at the time of hospitalisation as the criterion for the selection of relevant records in hospital discharge data, enabling a 30% reduction in false positive rates for the notification of breast cancer. Nevertheless, sensitivity, already poor, dropped from 69% to 64% [14] so that it was impossible to envisage exhaustive notification solely on the basis of hospital discharge data.

To our knowledge, no research has been conducted on the implementation of a selection among the variety of information available for a given individual from different data sources. Our hypothesis is that adding a selection step could reduce the noise resulting from incorrect or imprecise coding, and would enable the extraction of relevant information upstream from case notification.

The aim of this study was therefore to develop and evaluate an algorithm selecting relevant records for the notification of incident cases of cancer according to the information available for each individual.

2. Background

2.1 Registration Rules for Multiple Primary Cancers

Figure 1 presents the structure of the ICD-O-3 classification. The ICD-O-3 is a multi-axial classification used in cancer registries in order to record the anatomical site (topography) and the histology of a neoplasm. The histology axis is coded on five digit. The first four digits correspond to the morphology (histological description) and the fifth digit indicates whether a tumour is malignant, benign, in situ, or uncertain (whether benign or malignant).

For the purpose of defining multiple tumours, groups of topography codes considered as a single site and groups of morphology codes considered as a single neoplasm are defined [3]:

- IARC topography (Topographic axis). This level aggregates the topography codes in 54 target classes.
- IARC morphology ( Morphological axis). This level aggregates the morphology codes (regardless of behaviour) in 17 target classes. It corresponds to an adaptation of the morphology groups defined by Berg [15].

If the morphological diagnoses fall into one category of the 17 target classes, and arise in the same primary site, they are considered to be the same tumour for the purpose of counting multiple primaries. For paired organs such as breast, the same rules are applied so that bilateral breast cancers are counted as one tumour even if asynchronous. If the morphological diagnoses fall into two or more of the 17 target classes, even if they concern the same site, the morphology is considered to be different, and two or more cases should be counted. For certain tumour morphologies (Kaposi sarcoma and tumour of the haematopoietic system), a single tumour is registered, independently from topography.

![Figure 1](https://example.com/image.png)  
**Figure 1** Structure of the ICD-O-3 code (with example of a prostate adenocarcinoma)
We used this granularity given as the recommended level for multiple primary cancers as the target classes to be identified by the notification algorithm.

### 2.2 Validation Process in the General Cancer Registry of Poitou-Charentes

The collection and analysis of medical data by this cancer registry received the approval of the French regulatory authorities. In compliance with national and international recommendations, since January 1, 2008 the general cancer registry of the Poitou-Charentes region (western France) has included any incident case of malignant invasive tumour (haematological malignancy and solid tumours not including baso-cellular skin carcinomas), in situ tumour, borderline ovarian tumour, and tumour of the brain or the bladder that are benign or where evolution is unpredictable, involving a subject regularly residing in the Poitou-Charentes region at the time of diagnosis. The general cancer registry of Poitou-Charentes is qualified by the French NCR since January 2013 based on the 2008 and 2009 registered data.

> Figure 2 presents the validation process in the Poitou-Charentes cancer registry that was used in 2008 for notification, registration and validation of incident
Figure 3  Example of execution of the selection algorithm for a given individual. In this example, all data are related to a unique individual.

Prepared data

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>PROCEDURES</th>
<th>TOPO</th>
<th>IARC TOPO</th>
<th>MORPHO</th>
<th>IARC MORPHO</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td></td>
<td>C199</td>
<td>Rectum</td>
<td>80003</td>
<td></td>
<td>15/06/2008</td>
</tr>
<tr>
<td>AP</td>
<td>Surgery</td>
<td>C209</td>
<td>Rectum</td>
<td>81443</td>
<td>Adenocarcinomas</td>
<td>15/06/2008</td>
</tr>
<tr>
<td>HD</td>
<td>Radiation</td>
<td>C189</td>
<td>Colon</td>
<td>80003</td>
<td></td>
<td>15/06/2008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10/07/2008</td>
</tr>
<tr>
<td>HD</td>
<td>Radiation</td>
<td>C189</td>
<td>Colon</td>
<td>80003</td>
<td></td>
<td>10/08/2008</td>
</tr>
<tr>
<td>IS</td>
<td></td>
<td>C509</td>
<td>Breast</td>
<td>80003</td>
<td></td>
<td>20/12/2008</td>
</tr>
<tr>
<td>AP</td>
<td>Surgery</td>
<td>C509</td>
<td>Breast</td>
<td>85003</td>
<td>Adenocarcinomas</td>
<td>20/12/2008</td>
</tr>
<tr>
<td>HD</td>
<td>Radiation</td>
<td>C502</td>
<td>Breast</td>
<td>80003</td>
<td></td>
<td>17/01/2009</td>
</tr>
<tr>
<td>AP</td>
<td>Surgery</td>
<td>C509</td>
<td>Breast</td>
<td>85003</td>
<td>Adenocarcinomas</td>
<td>17/01/2009</td>
</tr>
<tr>
<td>HD</td>
<td>Radiation</td>
<td>C509</td>
<td>Breast</td>
<td>80003</td>
<td></td>
<td>02/03/2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>Radiation</td>
<td>C509</td>
<td>Breast</td>
<td>80003</td>
<td></td>
<td>27/03/2009</td>
</tr>
</tbody>
</table>

Topography selection

Selected topographies

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>PROCEDURES</th>
<th>TOPO</th>
<th>IARC TOPO</th>
<th>MORPHO</th>
</tr>
</thead>
<tbody>
<tr>
<td>HD</td>
<td>Surgery</td>
<td>C189</td>
<td>Colon</td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>Surgery</td>
<td>C502</td>
<td>Breast</td>
<td></td>
</tr>
</tbody>
</table>

Morphology selection

Selected morphologies

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>TOPO</th>
<th>IARC TOPO</th>
<th>MORPHO</th>
<th>IARC MORPHO</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>C209</td>
<td>Rectum</td>
<td>81443</td>
<td>Adenocarcinomas</td>
</tr>
<tr>
<td>AP</td>
<td>C509</td>
<td>Breast</td>
<td>85003</td>
<td>Adenocarcinomas</td>
</tr>
</tbody>
</table>

Tumour construction

Composite records

<table>
<thead>
<tr>
<th>TOPO</th>
<th>IARC TOPO</th>
<th>MORPHO</th>
<th>IARC MORPHO</th>
</tr>
</thead>
<tbody>
<tr>
<td>C189</td>
<td>Colon</td>
<td>81443</td>
<td>Adenocarcinomas</td>
</tr>
<tr>
<td>C502</td>
<td>Breast</td>
<td>85003</td>
<td>Adenocarcinomas</td>
</tr>
<tr>
<td>C502</td>
<td>Breast</td>
<td>85003</td>
<td>Adenocarcinomas</td>
</tr>
</tbody>
</table>

cases of cancer. The registry routinely collects four types of data sources using various terminologies to describe diagnosis:

- Anatomical pathology (AP) data that includes free-text reports related to one or several ADICAP diagnostic codes (Association pour le Développement de l’Informatique en Cytologie et en Anatomie Pathologique – French classification of lesions with topographical and histological axis).
- Hospital discharge (HD) data recorded in the French medical information program that includes ICD-10 diagnostic codes and CCAM medical procedure coded fields (Classification Commune des Actes Médicaux – the health insurance classification).
- Reimbursement of a dispensation for a cancer granted by the French healthcare insurance service (IS) that includes ICD-10 diagnostic codes.
- Data from the continuous survey in cancer care centres ("les Centres de Lutte contre le Cancer") that includes ICD-O-3 tumour codes. The continuous survey is a system of data collection in oncology promoted in France in 1975 by the National Federation of Cancer Care Centres. The continuous survey allows the identification of topographic and histological diagnosis of tumours, their initial extension, as well as thera-
peutic and evolutionary data for all cases in a given cancer care centre.

Data extractions include patient identity (name, surname, birthdate ...). When loading data in the registry's information system, patient's identities are integrated in an identity server which identifies with a semi-automated process data that are related to a single patient. The patient's identification process, based on computerized record linkage [5], allows automated linkage (based on deterministic rules), ambiguity detection, duplicate search and manual patient grouping or separation. This process allows all eligible records from data sources integrated in the information system to be related to single patients.

In order to permit a semantic integration, all diagnostic codes from data sources are transformed into ICD-O-3 using a terminology server. Then, the notification algorithm processes every record prepared by the semantic integration process and uses ICD-O-3 diagnostic codes to determine, for each patient, tumours that should be notified to cancer registry staff regarding registration rules for multiple primary cancers [3] and already validated tumours.

All tumour records automatically created by the notification algorithm contain related information including date of diagnosis, ICD-O-3 topography, ICD-O-3 morphology, basis of diagnosis and a zero level of validity. Then each case is manually checked by registry staff by visual inspection of information sources, assessing the need of a patient's medical record investigation for recording the tumour.

Only tumours that have never been manually modified can be updated through the notification process. The notification algorithm is executed each time new records are integrated from a data source and uses the entire information available for individuals that have at least one new record.

### 3. Materials and Methods

#### 3.1 Selection Algorithm for Relevant Records

Our approach was to add a selection step of relevant records between the semantic integration step and the notification algorithm (Figure 2). This approach was guided by the knowledge of data sources that the registry staff has acquired during manual data validation.

Throughout patient care several practitioners from different specialities and different structures can have occasion to implement coding procedures resulting in multiple tumour occurrence in the data extracted by the registry for secondary use. The accuracy of this coding, whether in terms of topography or morphology of the tumour, may be highly dependent on the way in which the practitioner approaches the case. We therefore set out some simple hypotheses concerning the choice of records liable to be used for notifying tumours:

- It is the clinician who is the best suited to notifying tumoral topography, in particular the clinician who performed the biopsy or the surgery. Indeed, the surgeon knows exactly the location of the tumour and is used to describe it.
- It is the pathologist who is best suited to notifying tumoral morphology and behaviour. Indeed, the pathologist knows exactly the histological type of the tumour and is used to describe it.
- The existence of multiple primary sites in an individual is not the norm, and notification requires an adequate level of proof.
- The remaining information should be taken into account when the data for a given individual is incomplete. This rule is essential in order to prevent from non-notification when some pertinent data are available for an individual.

Based on these assumptions the algorithm selects the best information at the patient level according to the data source providing the diagnostic code for tumoral topography and tumoral morphology independently.
3.1.1 Selection of Qualifying Topographies

Figure 3 presents an example of execution of the selection algorithm for a given individual. The selection algorithm is based upon the following hierarchy:
1. Cancer care centres’ data or Hospital discharge data → Diagnosis associated with surgical procedure or biopsy
2. Hospital discharge data → Diagnosis associated with lymph node dissection
3. Hospital discharge data → Diagnosis with no associated tracer procedure
4. Anatomical pathology data
5. Other diagnostic codes

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>PROCEDURES</th>
<th>TOPO</th>
<th>IARC TOPO</th>
<th>MORPHO</th>
<th>IARC MORPHO</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>IS</td>
<td>C199</td>
<td>Rectum</td>
<td>80003</td>
<td>Unspecified</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AP</td>
<td>C209</td>
<td>Rectum</td>
<td>81443</td>
<td>Adenocarcinomas</td>
</tr>
<tr>
<td></td>
<td>Surgery</td>
<td>HD</td>
<td>C189</td>
<td>Colon</td>
<td>80003</td>
<td>Unspecified</td>
</tr>
<tr>
<td></td>
<td>Radiation</td>
<td>HD</td>
<td>C189</td>
<td>Colon</td>
<td>80003</td>
<td>Unspecified</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Radiation</td>
<td>HD</td>
<td>C189</td>
<td>Colon</td>
<td>80003</td>
<td>Unspecified</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IS</td>
<td>C509</td>
<td>Breast</td>
<td>80003</td>
<td>Unspecified</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AP</td>
<td>C509</td>
<td>Breast</td>
<td>85003</td>
<td>Adenocarcinomas</td>
</tr>
<tr>
<td></td>
<td>Surgery</td>
<td>HD</td>
<td>C502</td>
<td>Breast</td>
<td>80003</td>
<td>Unspecified</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AP</td>
<td>C509</td>
<td>Breast</td>
<td>85003</td>
<td>Adenocarcinomas</td>
</tr>
<tr>
<td></td>
<td>Radiation</td>
<td>HD</td>
<td>C509</td>
<td>Breast</td>
<td>80003</td>
<td>Unspecified</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Radiation</td>
<td>HD</td>
<td>C509</td>
<td>Breast</td>
<td>80003</td>
<td>Unspecified</td>
</tr>
</tbody>
</table>

Figure 5 Example of execution of the notification algorithm with and without the selection step. Composite records derived from the selection algorithm are grouped to eliminate duplicate records. Notification algorithm without selection produced three tumours in application of registration rules for multiple primary cancers. The addition of the selection step produced only two tumours by excluding “rectum topography” reducing the noise generated by imprecise coding.
The aim of this hierarchy is to use, if available, the diagnostic code that was produced by the surgeon (or his staff) and the cancer care data. Given the data available for an individual, the algorithm retains only topographies derived from the source records corresponding to the highest hierarchical level available for the individual (level 1 corresponds to the highest possible level). Thus the different topographies derived from records at a lower level are ignored. For instance, in Figure 3 only two records associated with surgery are selected for topography notification. If the patient did not have surgery records then hospital discharge data with radiation procedure would have been selected for topography notification.

3.1.2 Selection of Qualifying Morphologies

The selection of morphologies is performed according to the same principle as for the topography, using the following hierarchy:

1. Cancer care centres’ data or Anatomical pathology data
2. Hospital discharge data ➔ Diagnosis associated with surgical procedure or biopsy
3. Hospital discharge data ➔ Diagnosis associated with lymph node dissection procedure
4. Hospital discharge data ➔ Diagnosis with no associated tracer procedure
5. Other diagnostic codes

The aim of this hierarchy is to use, if available, the diagnostic code that was produced by the pathologist (or his staff) and the cancer care data. For instance, in Figure 3 only three records from anatomical pathology data are selected for morphological notification.

3.1.3 Construction of Tumours

From the topographies and morphologies selected independently, the records are reconstructed for notification. Morphologies and topographies are linked one to the other, using following rules:

- If, for a given individual, there is strictly only one topography and only one morphology
  - The notified tumour is constructed by associating topography and morphology
- Otherwise
  - If a selected topography is available in selected records for morphologies
    - A notified tumour is constructed by associating this topography and this morphology
    - This topography and this morphology are excluded for further processing
    - The algorithm restarts from the beginning with remaining records
  - Among the remaining records
    - All possible tumours are constructed by connecting the available topographies and morphologies.

For instance, in Figure 3 “breast topography” is available in the selected morphology records describing “adenocarcinoma”. As a result the selected topography record (C502) is combined to the selected morphologies records (85003) to build two composite records. There is no more ambiguity among the remaining topography (C189) and morphology (81443). As a result, they are then combined in a unique composite record. In that example, the tumour construction, leads to three records available for the notification algorithm.

Once this processing is complete, a record is a composite of topography and morphology derived from different data sources. The records derived from the selection algorithm are then formatted to be fed directly into the registry notification algorithm.

3.2 Evaluation

The analysis was performed on the manually validated data derived from a full year registration (2008). Validated data comprise all the records present in the source data and attached to a tumour for which at least the topography, the morphology, the incidence date and the diagnostic basis have been finally validated by one of the physicians in the registry. In 2008, the registration process was as described in Figure 2. Every validated tumour had been notified by the automated notification process (without selection) and manually validated by the registry staff.

3.2.1 Evaluation Strategy for the Performance of the Algorithms

Target tumours are defined as the set of tumours validated by the physicians in the registry and included in the evaluation. Notified tumours are defined as the tumours that are automatically produced by an algorithm from source data. Concordant tumours are the notified tumours for which IARC topography, IARC morphology and tumoral behaviour correspond to the target tumour validated for an individual.

The evaluation consisted in comparing performances of an approach that does not include selection of relevant records (non-selective approach) with an approach including selection (selective approach). In either case tumours were notified from the data available, using the notification algorithm that was in routine use in the Poitou-Charentes cancer registry in 2008.

For each approach (selective and non-selective) the evaluation was performed by comparing the “target tumours” with the “notified tumours” for each patient (Figure 4). Figure 5 presents the execution of the two methods for the example presented in Figure 3. Composite records derived from the selection algorithm are grouped to eliminate duplicate records. Notification algorithm without selection produced three tumours in application of registration rules for multiple primary cancers. Elsewhere notification algorithm with a step selection produced only two tumours reducing the noise generated by imprecise coding. This evaluation involves the production of the objective measures that are presented in the next paragraph.

3.2.2 Evaluation Measures: Recall, Precision, F-measure

For each individual a sparse binary matrix was constructed representing presence or absence of each type of tumour. From this
We compared performances obtained using the notification algorithm with its own, among the 16,014 tumours notified by the selective approach, 8996 (72%) were concordant with the target tumours. Among the errors identified, 1720 (49%) were due to the production of tumours for which the topography was correct but the morphology imprecise (rather than incorrect). An anatomical pathology record was available for 1283 of these tumours encoded with an imprecise (rather than incorrect) morphology, or 37% of the discordant instances identified overall. Regarding the results for the notification algorithm on its own, among the 16,014 tumours notified, 7684 (48%) were concordant.

Table 1 presents the types of data available per individual. Almost 90% of the individuals had an anatomical pathology record available, and almost 69% an eligible diagnosis associated with a tracer procedure.

The validated tumours corresponded to 48 different topographies and 17 different morphologies, or 220 different types of tumour.

### 4.2 Algorithm Performances

#### 4.2.1 Global Performances

The selective approach notified 12,493 tumours as compared to 16,014 from the notification algorithm on its own. Among the 12,493 tumours notified by the selective approach, 8996 (72%) were concordant with the target tumours. Among the errors identified, 1720 (49%) were due to the production of tumours for which the topography was correct but the morphology imprecise (rather than incorrect). An anatomical pathology record was available for 1283 of these tumours encoded with an imprecise morphology, or 37% of the discordant instances identified overall. Regarding the results for the notification algorithm on its own, among the 16,014 tumours notified, 7684 (48%) were concordant.

Table 2 presents the performances of the two approaches for the notification of the different coding axes (IARC topography, IARC morphology, behaviour), and

#### Table 1: Type of data available per individual

<table>
<thead>
<tr>
<th>Type of data</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomical pathology data</td>
<td>10,757</td>
<td>89.9</td>
</tr>
<tr>
<td>Diagnosis with tracer procedure</td>
<td>8254</td>
<td>68.9</td>
</tr>
<tr>
<td>Healthcare insurance service</td>
<td>7241</td>
<td>60.5</td>
</tr>
<tr>
<td>Diagnosis with no tracer procedure</td>
<td>1445</td>
<td>12.1</td>
</tr>
<tr>
<td>Cancer care centres' data</td>
<td>664</td>
<td>5.5</td>
</tr>
</tbody>
</table>

For instance, for the analysis of the notification of complete tumours, an individual with the source type “Diagnosis with tracer procedure and anatomical pathology data” has no record of the type “Cancer care centres’ data”.

#### 4.3.3 Analyses Performed

1. We compared performances obtained by the use of the selection step along with the notification algorithm with performances obtained using the notification algorithm on its own, collating the results for “complete” tumours and the three axes defining a tumour (IARC topography, IARC morphology, behaviour). In order to produce synthetic global measures from measures calculated locally for each type of tumour, we used the micro-averaging method [16]: recall and precision were weighted by the corresponding number of individuals.

2. We compared the performances of the two approaches according to the types of data available for each individual. For this comparison, a single type of data was allocated hierarchically to each individual. Different hierarchies were used according to the axis studied:

- Complete tumour
  1. Cancer care centres’ data
  2. Diagnosis with tracer procedure and anatomical pathology data
  3. No tracer procedure and anatomical pathology data
  4. Other

- Topography
  1. Cancer care centres’ data
  2. Diagnosis with tracer procedure
  3. Diagnosis without tracer procedure
  4. Other

- Morphology
  1. Cancer care centres’ data
  2. Anatomical pathology data
  3. Other

For instance, for the analysis of the notification of complete tumours, an individual with the source type “Diagnosis with tracer procedure and anatomical pathology data” has no record of the type “Cancer care centres’ data”.

### 4.4 Results

#### 4.4.1 Data Description

For the evaluation, 11,971 individuals with 12,346 manually validated tumours were included. Among these individuals, 11,604 had a single tumour (97%) and 367 had multiple primary sites (3%). The mean number of sources per individual (among hospital discharge data, cancer care centres’ data, healthcare insurance data and anatomical pathology data) was 2.4. The mean number of records per individual was 11.2 (minimum 1 record, maximum 129 records). To execute the notification algorithms the following were used:

- Complete tumour
  1. Cancer care centres’ data
  2. Diagnosis with tracer procedure and anatomical pathology data
  3. No tracer procedure and anatomical pathology data
  4. Other

- Topography
  1. Cancer care centres’ data
  2. Diagnosis with tracer procedure
  3. Diagnosis without tracer procedure
  4. Other

- Morphology
  1. Cancer care centres’ data
  2. Anatomical pathology data
  3. Other

- Granularity

<table>
<thead>
<tr>
<th>Granularity</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Select</td>
<td>N-select</td>
<td>Select</td>
</tr>
<tr>
<td>Tumour topography</td>
<td>0.927</td>
<td>0.965</td>
<td>0.925</td>
</tr>
<tr>
<td>Tumour morphology</td>
<td>0.805</td>
<td>0.809</td>
<td>0.804</td>
</tr>
<tr>
<td>Tumour behaviour</td>
<td>0.894</td>
<td>0.826</td>
<td>0.909</td>
</tr>
<tr>
<td>Complete tumour</td>
<td>0.729</td>
<td>0.622</td>
<td>0.721</td>
</tr>
</tbody>
</table>

a With selection algorithm, b without selection algorithm
globally for complete tumours. The selective approach demonstrates better performances for the notification of complete tumours (0.72 and 0.54 f-measures for topography and morphology respectively). This improvement holds for all the tumour components (IARC topography, IARC morphology, behaviour).

Regarding topography, the gain in performance is in terms of precision at the expense of recall, but with an improvement in the F-measure. For IARC morphology, recall is altered only slightly, and there is a clear gain in precision. For behaviour, both recall and precision are improved by the new algorithm. The morphological axis is the one that presents the least satisfactory performances.

### 4.2.2 Performances According to the Type of Data Available

Table 3 presents the performances of the algorithms for the different coding axes and for complete tumours according to sources available for the different individuals. The selective approach shows better performances for the notification of complete tumours, with an improvement in recall and precision whatever the sources available.

<table>
<thead>
<tr>
<th>Type of data</th>
<th>N</th>
<th>Recall Selectb</th>
<th>Recall N-selectc</th>
<th>Precision Selectb</th>
<th>Precision N-selectc</th>
<th>F-Measure Selectb</th>
<th>F-Measure N-selectc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notification of complete tumour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer care centre’s data</td>
<td>664</td>
<td>0.904</td>
<td>0.836</td>
<td>0.785</td>
<td>0.570</td>
<td>0.840</td>
<td>0.677</td>
</tr>
<tr>
<td>Diagnosis with tracer procedure and anatomical pathology data</td>
<td>7380</td>
<td>0.752</td>
<td>0.632</td>
<td>0.742</td>
<td>0.483</td>
<td>0.747</td>
<td>0.548</td>
</tr>
<tr>
<td>Diagnosis with no tracer procedure and anatomical pathology data</td>
<td>2798</td>
<td>0.713</td>
<td>0.593</td>
<td>0.735</td>
<td>0.471</td>
<td>0.724</td>
<td>0.525</td>
</tr>
<tr>
<td>Other</td>
<td>1129</td>
<td>0.504</td>
<td>0.504</td>
<td>0.501</td>
<td>0.414</td>
<td>0.502</td>
<td>0.454</td>
</tr>
<tr>
<td>Notification of tumour topography</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer care centre’s data</td>
<td>664</td>
<td>0.981</td>
<td>0.987</td>
<td>0.893</td>
<td>0.703</td>
<td>0.935</td>
<td>0.821</td>
</tr>
<tr>
<td>Diagnosis with tracer procedure</td>
<td>7725</td>
<td>0.948</td>
<td>0.985</td>
<td>0.940</td>
<td>0.764</td>
<td>0.944</td>
<td>0.861</td>
</tr>
<tr>
<td>Diagnosis with no tracer procedure</td>
<td>1549</td>
<td>0.878</td>
<td>0.941</td>
<td>0.873</td>
<td>0.643</td>
<td>0.876</td>
<td>0.764</td>
</tr>
<tr>
<td>Other</td>
<td>2033</td>
<td>0.868</td>
<td>0.898</td>
<td>0.918</td>
<td>0.835</td>
<td>0.892</td>
<td>0.865</td>
</tr>
<tr>
<td>Notification of tumour morphology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer care centre’s data</td>
<td>664</td>
<td>0.939</td>
<td>0.942</td>
<td>0.862</td>
<td>0.717</td>
<td>0.899</td>
<td>0.814</td>
</tr>
<tr>
<td>Anatomical pathology data</td>
<td>10178</td>
<td>0.825</td>
<td>0.830</td>
<td>0.826</td>
<td>0.697</td>
<td>0.825</td>
<td>0.758</td>
</tr>
<tr>
<td>Other</td>
<td>1129</td>
<td>0.548</td>
<td>0.543</td>
<td>0.557</td>
<td>0.487</td>
<td>0.553</td>
<td>0.514</td>
</tr>
</tbody>
</table>

* N is the number of individuals per class, b with selection algorithm, c without selection algorithm

Regarding IARC topography, a clear improvement is observed in the F-measure in presence of a diagnosis with tracer procedure, and a diagnosis without any procedure. This improvement results from a considerable increase in precision, alongside a slight decrease in recall. The selective approach exhibits better performances for the notification of IARC morphology when anatomical pathology records and cancer care centres’ records are available. This improvement is largely due to a gain in precision, without markedly affecting recall.

### 5. Discussion

Our results demonstrate a clear improvement in performance with the implementation of the selective approach. We have shown the advantage of not using all the records for the notification phase. An approach involving a hierarchized selection makes it possible to process each individual case in differentiated manner, adapting the selection to the information available. In a large number of cases this enables the reduction of noise resulting from imprecision or inaccuracies in coding procedures. The notification of false multiple primary sites was reduced, and as a result the number of tumours produced after application of the hierarchical selection algorithm was far more realistic than when there was no selection procedure.

The initial hypotheses concerning the quality of coding according to the practitioner appear justified. Cancer care centres’ data constitutes a very reliable source for notification, as the synthesis that is routinely produced by these centres is very close to the registry targets. However the absence of anti-cancer centres on the Poitou-Charentes region resulted in a small proportion of cases (5.5%) being detected via cancer care centres’ data. For the notification of tumoral topography, the strategy, where a tracer procedure exists, of removing all the other records without a tracer procedure enables a marked improvement in precision with no major change in recall. This shows that the topographies associated with these records are accurate, and that the loss of information as a result of this selection is small. For the notification of morphology, anatomical pathology data is an essential source when there is no cancer care centre’s data coding. Nevertheless, the use of complementary sources when the case file is incomplete enables a
degree of exhaustiveness to be preserved, minimising damage in terms of precision. There is discordance between the decrease in recall for the two axes, topography and morphology, and increase in recall for the complete tumour. This observation results from the construction by the selection algorithm of a composite record describing the tumour according to these two axes. In this composite record, morphology and topography can derive from different sources. If there is no selection algorithm, the notification algorithm creates two tumours. In Figure 5, without selection, two tumours are generated (colon-unspecified and rectum adenocarcinomas). If the target tumour is (colon adenocarcinomas), no tumour matches this target whereas the record generated with the selection algorithm lead to create the target tumour thus improving both recall and precision. The recall discordance observed clearly shows that although the selection procedure generates a loss of information for each axis, the synthesis is more relevant for the notification of a complete tumour.

In anatomical pathology data, there are non-coded records (the practitioner has not coded the tumour) or else the coding is not useable (coding irregularity). In this situation, our algorithm notifies an imprecise morphology coding. In the data analysed here, this type of error accounted for 37% of the discordant tumours observed in the selective approach. An earlier study implemented a text categorisation method which demonstrated very good performances in the case of single non-metastatic tumours, with an F-measure of 0.97 for IARC morphological coding [17]. Its integration upstream from the selection algorithm could automatically generate a morphological group for these records, and would probably improve the global notification performances.

In Italy, the Varese Cancer Registry estimated the proportion of wrongly-coded topographies from their data sources to be 1.6% on a full year registration [8, 11]. Another evaluation performed on 1539 files in the Venetian Tumour Registry showed 85.7% concordance on IARC topography and IARC morphology and only 39 false positives (2.5%) [12]. These evaluations concerned cases that had several concordant sources leading to describe only one single tumour for a given patient and were then eligible to be registered automatically in these registries. These tumours were representing about 60% of the total number of registered tumours in 1997. These results are difficult to compare with ours because we used a complete year of registration. Our aim was to deal with coding inconsistencies that led to non-concordant coding for a single individual. These complex cases were excluded from evaluations performed in Italy. The Ontario Cancer Registry estimated the proportion of error to be 6.7% for notification of IARC topographies but the analysis was restricted to cases in which only a single primary had been registered for the patient [5, 18] corresponding to 91% of all registered tumours in this registry. To our knowledge, no evaluation is available about patient with automated multiple primary cancers. Our algorithm produces 7.5% errors for IARC topographic groups and 72% of complete tumours notified were concordant. Our evaluations were performed on a full year registration including multiple primary cancers.

The method developed in Italy in order to automatically register cases is based on diagnostic codes: “To choose the best code a hierarchy was defined whereby a specific site (ICD-9) code takes precedence over a generic site code, which in turn takes precedence over a metastatic or ill-defined site code, or a code indicating secondary unspecified lymph node involvement” [8]. The registration algorithm developed in Italy aims at determining the concordance between diagnostic codes available through multiple data sources. The selection algorithm we developed is an additional step that uses information associated to diagnostic codes (such as source or associated procedures) rather than diagnostic codes themselves. By selecting, for a given individual, relevant records (and excluding non-relevant record) before evaluating diagnostic codes concordance, the selection algorithm reduces the effect of coding inconsistencies. These inconsistencies might represent a significant part of the cases that cannot be registered automatically due to non-concordant diagnostic codes. Tognazzo et al. [13] proposed a probabilistic approach in order to find new cancer cases. Their evaluations showed that the source of a discordant diagnostic code was significantly influencing both random forest and multi-logit model. The deterministic selection process that we developed shows that information associated with diagnostic codes (such as data source and procedures) plays a significant role in new cases definition.

Notification of tumours as accurately as possible is a genuine asset for the manual validation phase. By decreasing the number of alterations required during validation as far as possible, while at the same time allowing the operator access to all the non-selected information, the approach necessarily improves the efficiency and the quality of the registration procedures.

The year of incidence was not included in the definition of the tumour. One of the main tasks in future work will be, from the data available, to detect situations in which a case is likely to be previous to the date of the first notification by one of the sources (delayed relapse, late instatement of care, therapeutic abstention). As the rules we applied were based on the registry staff knowledge of data sources, we did not evaluate other hierarchical selection. The result of this study suggests that the context of a diagnostic code plays a significant role. The Poitou-Charentes cancer registry information system provides a manually validated relation between tumours and data sources. This data set will be used to build learning and test sets in order to develop and evaluate probabilistic approaches with development of context encoding.

The ongoing work on improvement of the quality of data in the information system of the Poitou-Charentes cancer registry has led to the development of a selection algorithm for relevant records. The present work has made it possible to demonstrate the value of a selective approach compared to an exhaustive approach in the definition of tumours to be notified. The use of only part of the information available for notification, via the hierarchical selection of qualifying records, enables the removal of a large part of the noise generated by imprecise or incorrect records. In a multi-source system, it is coherent to use a typology of information to determine its qualifying relevance. It should however be noted that this approach is envisaged solely in the context of a manual validation of
case files by an operator who has access to all the information (not merely the hierarchized selection). Although the approach consisting in selecting information for notification might be viewed as risky, it can nevertheless be considered coherent and amply justified when implemented in association with the development of an exhaustive summary of care trajectory, enabling the operator to have access to all the available information for case validation.

Acknowledgements

The authors would like to thank the medical records assistants, Nicolas Mériau, Sébastien Orazio and Soizic Lellouch of the Registre des cancers de Poitou-Charentes for their helpful contribution to the implementation of this research. We also wish to thank Angela Swaine Verdier for the translation of the manuscript.

References