



# Représentation des diagnostics de tumeurs utilisés par les registres des cancers par l'intermédiaire du NCI thesaurus : évaluation de cette terminologie comme outil pour l'utilisation conjointe de données en cancérologie

Bérénice Bréchat-Huet

## ► To cite this version:

Bérénice Bréchat-Huet. Représentation des diagnostics de tumeurs utilisés par les registres des cancers par l'intermédiaire du NCI thesaurus : évaluation de cette terminologie comme outil pour l'utilisation conjointe de données en cancérologie. Médecine humaine et pathologie. 2015. dumas-01206467

**HAL Id: dumas-01206467**

**<https://dumas.ccsd.cnrs.fr/dumas-01206467>**

Submitted on 29 Sep 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**UNIVERSITE DE BORDEAUX**  
**U.F.R. DES SCIENCES MEDICALES**

Année 2015

Thèse n° 3095

Thèse pour l'obtention du  
**DIPLOME D'ETAT DE DOCTEUR EN MEDECINE**

Présentée et soutenue publiquement le 17 septembre 2015

**Par Bérénice BRECHAT-HUET**

Née le 14 juillet 1982 à Sion (C.H.)

**Représentation des diagnostics de tumeurs utilisés par les registres des cancers  
par l'intermédiaire du NCI thesaurus : évaluation de cette terminologie comme  
outil pour l'utilisation conjointe de données en cancérologie.**

|                                       |            |
|---------------------------------------|------------|
| Professeur Simone Mathoulin-Pelissier | Président  |
| Professeur Anita Burgun               | Rapporteur |
| Docteur Frantz Thiessard              | Juge       |
| Docteur Renaud Besselère              | Juge       |
| Docteur Fleur Mougin                  | Juge       |
| Docteur Vianney Jouhet                | Directeur  |



## Remerciements

*Au Professeur Simone Mathoulin-Pelissier,*

pour m'avoir fait l'honneur de présider le jury de cette thèse, et pour tout ce que j'ai pu apprendre, professionnellement et humainement, à ses côtés.

*Au Docteur Frantz Thiessard,*

pour son implication dans l'enseignement de notre discipline qu'est l'informatique médicale et pour sa profonde gentillesse.

*Au Docteur Renaud Besselère,*

pour nos discussions si enrichissantes et la passion qu'il m'a communiquée. Merci de faire partie de ce jury.

*Au Docteur Fleur Mougin,*

pour tout le calme et la sagesse apportés à ce projet, pour ses nombreuses relectures et sa bonne humeur.

*Au Docteur Vianney Jouhet,*

Pour m'avoir proposé ce sujet et guidé dans ce travail. Merci d'avoir apporté un écho à mon obstination, qu'il provoque de belles choses...

*Au Professeur Anita Burgun,*

pour l'intérêt témoigné envers ce travail de thèse en acceptant d'en être le rapporteur et pour son apport à la Santé Publique en général.



Je remercie toutes les personnes qui m'ont apporté leur aide dans la réalisation de cette thèse.

Merci à l'équipe du registre, qui a partagé avec moi tant de moments importants. Merci aux membres de l'ERIAS de m'avoir accompagné pendant ces dernières années.

J'adresse une pensée particulière à Marianne Savès, pour sa bienveillance sur mon parcours.

Merci aux autres internes de santé publique, en particulier Pauline, Isabelle et Florence, ISP de cœur, pour leur amitié et leur soutien.

Un grand merci à tous mes proches. J'espère qu'ils savent à quel point ils sont importants.

A Fanny, Gwenaëlle et Claire pour leur amitié et leur précieux savoir.

A Marie, Clémence et Aurélie pour avoir partagé tout cela avec moi.

Au P66.

Et évidemment, à Sylvain.

A mes parents, qui comptent tellement, et à Jean-Baptiste et Léonie, sans qui rien n'a de sens.



# TABLE DES MATIERES

|                                                                                                   |           |
|---------------------------------------------------------------------------------------------------|-----------|
| <b>1. DEFINITIONS .....</b>                                                                       | <b>3</b>  |
| <b>2. CONTEXTE.....</b>                                                                           | <b>5</b>  |
| <b>2.1. Terminologies biomédicales multiples et interopérabilité.....</b>                         | <b>5</b>  |
| <b>2.2. Utilisation conjointe des données issues de différentes sources en cancérologie .....</b> | <b>5</b>  |
| <b>2.3. Choix du NCIt comme base de la terminologie cible .....</b>                               | <b>7</b>  |
| 2.3.1. Notions de pré et post-coordination de concepts.....                                       | 7         |
| 2.3.2. Structure du NCIt .....                                                                    | 8         |
| <b>2.4. Choix de la CIM-O3 comme première terminologie à mapper .....</b>                         | <b>9</b>  |
| 2.4.1. CIM-O3 : une terminologie diagnostique de référence.....                                   | 9         |
| 2.4.2. Structure de la CIM-O3 .....                                                               | 10        |
| <b>2.5. Mise à l'épreuve du mapping.....</b>                                                      | <b>11</b> |
| 2.5.1. Confrontation à des données réelles.....                                                   | 11        |
| 2.5.2. Registre général des cancers de la Gironde.....                                            | 11        |
| 2.5.2.1. Sources d'information du RGCG.....                                                       | 12        |
| 2.5.2.2. Critères d'inclusion des cas .....                                                       | 12        |
| 2.5.2.3. Codage du diagnostic .....                                                               | 13        |
| <b>2.6. Objectif .....</b>                                                                        | <b>13</b> |
| <b>3. MATERIEL ET METHODE .....</b>                                                               | <b>14</b> |
| <b>3.1. Mapping de la CIM-O3 vers le NCI thesaurus .....</b>                                      | <b>14</b> |
| 3.1.1. Etape 1 : recherche de mappings candidats.....                                             | 16        |
| 3.1.1.1. NCI metathesaurus.....                                                                   | 16        |
| 3.1.1.2. Mappings candidats .....                                                                 | 17        |
| 3.1.2. Etape 2 : évaluation des mappings candidats.....                                           | 18        |
| 3.1.3. Etape 3 : enrichissement manuel de la terminologie cible.....                              | 19        |
| <b>3.2. Représentations des diagnostics de tumeurs du RGCG au sein du NCIt .....</b>              | <b>19</b> |
| 3.2.1. Extraction des diagnostics de cancers codés dans la base du registre .....                 | 19        |
| 3.2.2. Mapping des topographies et morphologies CIM-O3 vers le NCIt .....                         | 20        |
| 3.2.3. Recherche de concepts pré-coordonnés .....                                                 | 21        |
| <b>3.3. Technologies utilisées.....</b>                                                           | <b>21</b> |
| <b>4. RESULTATS.....</b>                                                                          | <b>22</b> |
| <b>4.1. Mapping de la CIM-O3 vers le NCI thesaurus .....</b>                                      | <b>22</b> |
| <b>4.2. Représentation des diagnostics de tumeurs du RGCG au sein du NCIt.....</b>                | <b>23</b> |
| 4.2.1. Description des diagnostics de tumeurs codés dans le RGCG .....                            | 23        |
| 4.2.2. Mapping des diagnostics CIM-O3 vers des concepts NCIt simples.....                         | 24        |
| 4.2.2.1. Résultat du mapping .....                                                                | 24        |
| 4.2.2.2. Description des mappings manquants .....                                                 | 25        |
| 4.2.3. Diagnostics pré-coordonnés .....                                                           | 26        |

|                                                                                |           |
|--------------------------------------------------------------------------------|-----------|
| <b>5. DISCUSSION .....</b>                                                     | <b>28</b> |
| <b>5.1. Synthèse des résultats et analyse .....</b>                            | <b>28</b> |
| 5.1.1. Mapping de la CIM-O3 vers le NCIt .....                                 | 28        |
| 5.1.2. Représentation des diagnostics du RGCG.....                             | 29        |
| 5.1.2.1. Diagnostics du RGCG .....                                             | 29        |
| 5.1.2.2. Mapping des composantes topographie et morphologie vers le NCIt ..... | 29        |
| 5.1.2.3. Recherche des diagnostics pré-coordonnés.....                         | 31        |
| <b>5.2. Limites potentielles .....</b>                                         | <b>32</b> |
| 5.2.1. Variabilité inter-humaine .....                                         | 32        |
| 5.2.2. Choix de validation sur les termes préférés .....                       | 32        |
| 5.2.3. Utilisation de diagnostics issus d'un registre général.....             | 33        |
| <b>5.3. Perspectives .....</b>                                                 | <b>33</b> |
| <b>6. CONCLUSION .....</b>                                                     | <b>34</b> |
| <b>BIBLIOGRAPHIE.....</b>                                                      | <b>35</b> |
| <b>LISTE DES ANNEXES.....</b>                                                  | <b>38</b> |
| <b>LISTE DES FIGURES .....</b>                                                 | <b>39</b> |
| <b>LISTE DES TABLES.....</b>                                                   | <b>39</b> |
| <b>Abréviations .....</b>                                                      | <b>40</b> |
| <b>Convention d'écriture .....</b>                                             | <b>40</b> |

# 1. DEFINITIONS

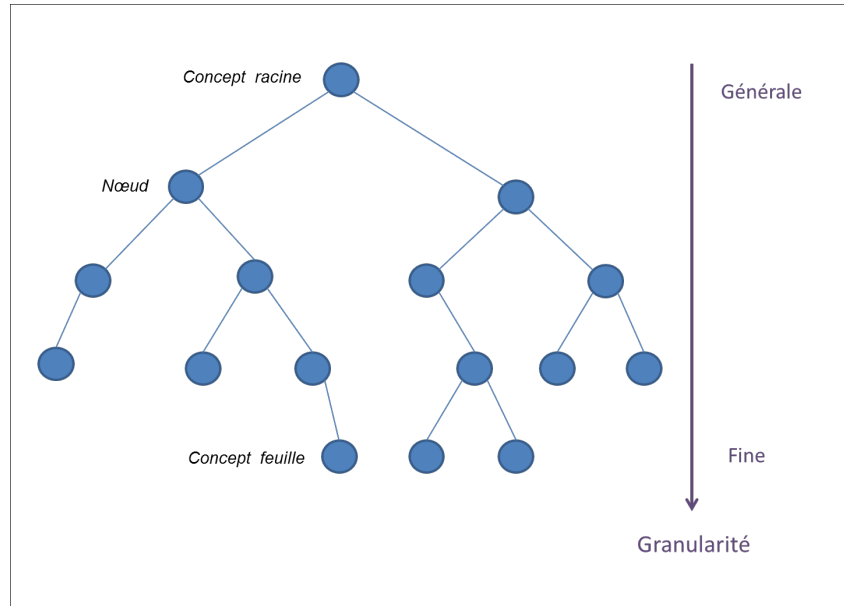
Une **terminologie** correspond à la représentation de l'ensemble des termes spécifiques d'un domaine particulier permettant de définir les objets ou concepts utilisés par ce domaine [1-2]. Ainsi, une terminologie recense les concepts d'un domaine, les termes qui désignent ces concepts ainsi que leurs définitions. Si la représentation de l'ensemble de ces termes est normalisée et qu'il existe une structuration des concepts les uns par rapport aux autres, on parle alors de **thesaurus**.

Un **concept** est une représentation générale et abstraite d'un objet, d'une situation ou d'un phénomène. Il peut être assimilé à la signification d'un terme. Le **code** correspond à une représentation alphanumérique du concept. En principe, un concept ne peut être représenté que par un seul code et chaque code équivaut à un seul concept. Ainsi, un code implique un concept et inversement.

La notion de **terme** correspond à l'expression du concept en langage naturel. On parle souvent de libellé, c'est-à-dire un mot ou un groupe de mots corrélés au concept, mis en relation avec un code dans une terminologie pour le définir. Un concept, donc un code, peut être représenté par plusieurs termes, et dans ce cas il est nécessaire de distinguer le « terme préféré », qui est le terme le plus souvent approprié dans le domaine spécifique pour représenter le concept, des termes synonymes.

Lorsque les concepts d'une terminologie sont organisés, il existe une structuration de ces derniers en fonction de leur précision, assimilable à des **hiérarchies** (*Figure 1*) : les concepts les plus génériques (également appelés concepts **racines**) donnent des branches avec des concepts enfants (également appelés nœuds) qui sont des concepts plus spécifiques. Les concepts enfants peuvent à leur tour donner plusieurs branches et ainsi de suite. Plus il y a de générations dans la hiérarchie, plus cette dernière est profonde (on dit que sa granularité est de plus en plus fine).

Chacun des éléments est subordonné à celui qu'il suit, chaque sous-concept est enfant d'un concept plus général (hormis la racine). Il peut ainsi exister des concepts enfants d'un concept, des concepts parents d'autres concepts, des concepts frères/sœurs lorsque deux concepts sont enfants du même concept. Le concept le plus fin de la branche, qui n'a pas de concept fils, est appelé concept terminal ou **feuille**.



**Figure 1 : Hiérarchie de concepts au sein d'une terminologie et granularité.**

Par ailleurs, il existe deux types de relations hiérarchiques : *is\_a* (*est\_un*) et *part\_of* (*partie\_de*). Ainsi, si le concept *cancer du sein* a une relation de type *is\_a* avec le concept *cancer*, cela signifie que le *cancer du sein* est un *cancer* (le concept de *cancer du sein* étant un élément plus précis que le concept de *cancer*). Si dans la famille des régions anatomiques, le concept *bras* est enfant de *membre supérieur* et qu'il existe des relations de type *part\_of* entre les générations de cette famille, cela signifie que la région *bras* est une partie de la région du *membre supérieur*, *bras* étant un concept plus fin que *membre supérieur*.

Il peut également exister d'autres types de relations, qui ne sont pas hiérarchiques, entre les concepts d'un même arbre, mais aussi entre les concepts d'arbres différents. Par exemple, dans une terminologie médicale, on peut mettre en relation un concept de la branche des maladies avec un concept de la branche des organes pour spécifier que cette maladie touche généralement cet organe, grâce à une relation transversale entre ces deux concepts.

## 2. CONTEXTE

### 2.1. Terminologies biomédicales multiples et interopérabilité

Les progrès exponentiels dans le domaine scientifique couplés au développement croissant des technologies informatiques sont aujourd'hui à l'origine d'une production massive d'information. Ainsi, les activités liées à la santé génèrent une grande quantité de données relatives aux soins des patients, qu'il faut stocker en vue d'une utilisation ultérieure éventuelle. Pour permettre un stockage sur un espace plus restreint, et surtout pour permettre une utilisation simplifiée de ces données, l'intérêt du codage de l'information est apparu avec une volonté de structurer cette information. Dans ce cadre, de nombreuses terminologies biomédicales ont été élaborées [3-7]. Malheureusement, le développement de ces terminologies s'est souvent fait localement, en fonction des besoins spécifiques des promoteurs et pour un usage précis. Il n'existait pas de consensus dans l'élaboration d'une terminologie, ni d'un point de vue structurel [8], ni d'un point de vue fonctionnel, ni d'un point de vue référentiel. Ainsi, au sein d'un même domaine, il peut exister des terminologies diverses destinées à des usages similaires mais qui codent l'information de façon très hétérogène, avec des concepts semblables mais des termes différents, et des structures de terminologies disparates [9,10]. Ceci pose de réels problèmes d'interopérabilité entre les terminologies et transposer de l'information codée grâce à une terminologie vers une autre terminologie est souvent synonyme de modification de l'information, voire de perte d'information. Mettre en relation deux concepts semblables peut paraître simple, mais lorsque ces concepts sont définis différemment, représentés par des termes distincts et mis en relation avec d'autres concepts de façon hétérogène, la tâche devient complexe. Le domaine de la recherche en informatique médicale cherche depuis plusieurs années à résoudre ce problème d'interopérabilité entre les terminologies d'un même domaine [11-14].

### 2.2. Utilisation conjointe des données issues de différentes sources en cancérologie

La cancérologie n'a pas échappé à la production désorganisée de terminologies spécifiques. Ainsi, pour coder un même diagnostic en cancérologie, on peut utiliser plusieurs terminologies en fonction de la spécialité, ou parfois au sein de la même spécialité. Les anatomo-pathologistes utilisent le plus souvent un thésaurus créé par l'Association pour le Développement de l'Informatique en Cytologie et en Anatomie Pathologique, appelé Thésaurus de la codification ADICAP [15]. En clinique, il est possible d'utiliser plusieurs terminologies différentes pour coder un diagnostic de cancer : certains utilisent l'une des versions de la

Classification Internationale des Maladies (CIM-9 ou CIM-10) [16], d'autres codent à l'aide du NCI thesaurus (NCIt) [17-18], développé par le *National Cancer Institute* américain. Ainsi, le Programme de Médicalisation du Système d'Information (PMSI) français utilise comme standard de codage des diagnostics la CIM-10. Les registres des cancers utilisent quant à eux la troisième version de la classification internationale des maladies pour l'oncologie (CIM-O3) [19], créée par l'Organisation Mondiale de la Santé (OMS), et cette terminologie est également celle utilisée pour le codage des diagnostics lors des réunions de concertation pluridisciplinaires en France. Toutes ces terminologies permettent de coder les mêmes diagnostics mais de façon différente, avec des codes ayant des structures différentes, et des terminologies ayant des architectures différentes. Ceci pose un réel problème lorsque l'on souhaite utiliser ensemble des données récoltées et codées par différentes structures. Par exemple si l'on souhaite mettre en relation des données d'une cohorte avec des statistiques nationales sur le cancer dont le diagnostic est exprimé dans une terminologie différente, ou si l'on souhaite constituer une population d'étude importante en réunissant des données semblables recueillies dans plusieurs pays, qui codent les diagnostics de façon différente, ou encore croiser des bases de données locales avec des bases nationales, par exemple les bases de l'assurance maladie. Il faut ainsi être en mesure de mettre en correspondance les diagnostics semblables entre eux, donc de pouvoir rattacher sous un même concept des codes différents, et ce sans modifier ou perdre de l'information. On parle de *mapping* entre deux concepts équivalents.

S'il existe déjà des mises en relation disponibles entre certaines terminologies diagnostiques, ces mappings sont rarement exhaustifs, peuvent comporter des erreurs, proposent souvent plusieurs correspondances possibles pour un même concept [21], et leur utilisation nécessitent une intervention humaine pour résoudre les erreurs engendrées. Ainsi, l'élaboration d'une terminologie commune qui permettrait de représenter chaque concept de chaque terminologie de manière unique et correcte est une solution de choix pour l'utilisation conjointe de données issues de différentes sources, mais qui reste néanmoins un challenge aujourd'hui. En effet, pour que cette terminologie commune remplisse sa fonction, il faut que chaque code de chaque terminologie pointe vers un et un seul code de la terminologie cible, et qu'il existe une équivalence entre les deux concepts *mappés*.

C'est dans la perspective d'élaborer une terminologie diagnostique cible de ce type en cancérologie que nous avons effectué une étude de faisabilité afin de mettre en évidence les problématiques potentielles rencontrées lors de la mise en correspondance des concepts et d'explorer les mappings existants pouvant servir de base à cette terminologie cible.



## 2.3. Choix du NCIt comme base de la terminologie cible

A l'issue de l'étude de faisabilité, le NCIt s'est avéré être une terminologie extrêmement intéressante comme base à l'élaboration de la terminologie cible, car sa structure permettait de répondre à une problématique majeure pour la mise en correspondance des terminologies diagnostiques en cancérologie : la coexistence de pré et post-coordination.

### 2.3.1. Notions de pré et post-coordination de concepts

Pour expliquer les notions de **pré et post-coordination**, considérons le codage des diagnostics en cancérologie. Un diagnostic correspond à une tumeur d'une morphologie donnée qui a une origine déterminée et qui concerne un organe donné (également appelé topographie). Par exemple, un adénocarcinome de la prostate est un diagnostic de tumeur précis, qui correspond à une tumeur dont le type morphologique est un adénocarcinome et qui se trouve dans la prostate.

Un élément majeur qui différencie la structure des terminologies diagnostiques en cancérologie est la possibilité de coder un même diagnostic de tumeur selon deux modalités différentes : de façon pré-coordonnée ou en post-coordonnant les concepts [22].

Ainsi, certaines terminologies représentent chaque tumeur particulière comme un concept à part entière, chaque code représente une morphologie précise dans un organe donné (par exemple, la CIM-10). Le concept d'adénocarcinome de la prostate sera codé dans son ensemble sous un code unique par ce type de terminologie. On peut qualifier ces terminologies de « pré-coordonnées » car l'élément de topographie et l'élément de morphologie qui composent une pathologie sont associés en amont du codage, lors de l'élaboration de la terminologie.

D'autres terminologies représentent séparément l'élément de topographie et l'élément de morphologie, chacun étant codé séparément (par exemple, la CIM-O3). Le cancer particulier est codé en combinant un code topographique et un code morphologique. Ainsi, la tumeur adénocarcinome de la prostate sera codée grâce à l'association entre le code du type histologique adénocarcinome et le code de l'organe prostate.

Avec les terminologies « pré-coordonnées », le codage est simplifié mais on ne peut pas coder une tumeur qui n'existe pas en tant que telle dans la terminologie. Il faut donc que ces terminologies soient exhaustives et régulièrement mises à jour. A contrario, il est possible de coder un large panel de tumeurs grâce à la post-coordination et être ainsi exhaustif. Cependant, l'utilisation des terminologies nécessitant une post-coordination peut entraîner des erreurs lors du codage puisqu'il est a priori possible d'associer chaque morphologie avec chaque topographie, même si dans la réalité, certaines associations sont impossibles. Pour pallier à ce type de problèmes, certaines terminologies précisent des règles d'association entre concepts, mais qui ne sont qu'indicatives.

Cette différence de structure des terminologies diagnostiques nécessite, pour mettre en correspondance les concepts de diagnostics, de pouvoir établir explicitement des équivalences entre un diagnostic et la combinaison d'une morphologie et d'une topographie. L'existence concomitante au sein du NCIt de ces deux types de structures de codage des diagnostics a motivé le choix d'utilisation de cette terminologie comme base pour l'élaboration de la terminologie cible.

### 2.3.2. Structure du NCIt

Le NCIt est une terminologie exprimée en langue anglaise qui couvre le vocabulaire utilisé en cancérologie au sens large, que ce soit en clinique, dans le domaine de la recherche, de la prévention, mais aussi dans le domaine administratif [17]. Cette terminologie est enrichie tous les mois par de nouveaux termes, répondant aux besoins des utilisateurs. Le NCIt comprend plus de 43 000 concepts représentant près de 10 000 cancers et pathologies associées, l'intégralité des concepts de la Federal Consolidated Health Informatics standard anatomy section (référence anglaise pour la représentation des concepts anatomiques), plus de 17 000 entités relatives aux traitements utilisés en cancérologie, des éléments de génétique, etc.

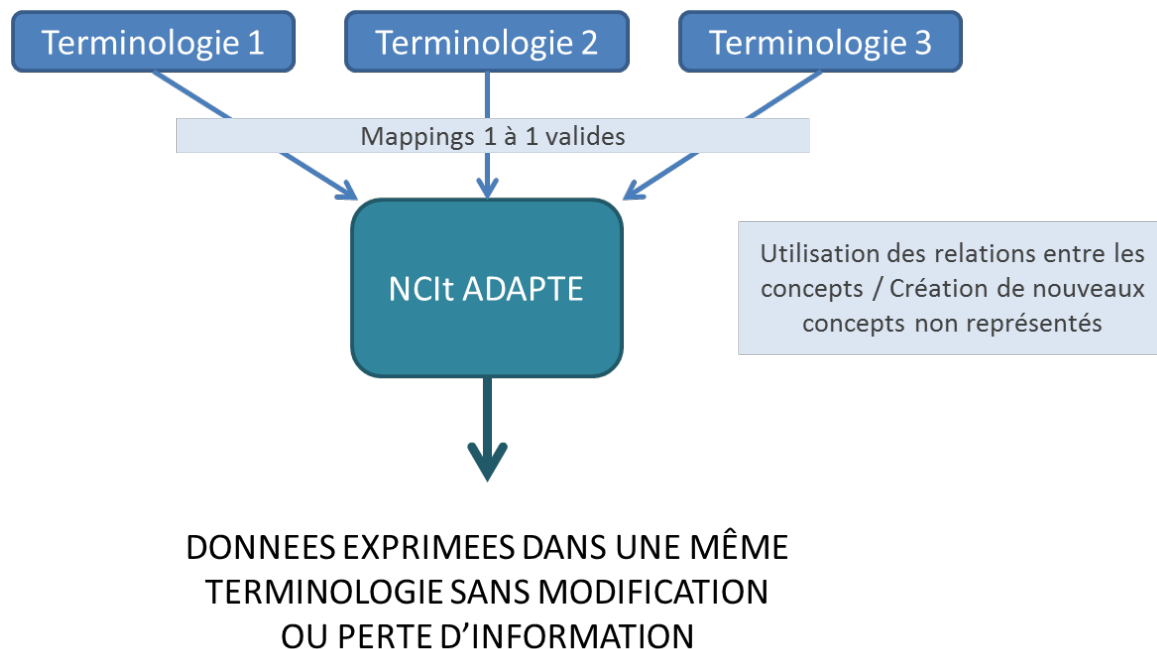
Ces concepts sont organisés selon 20 sous-domaines appelés *Kind* (listés en annexe 1) tels que : « Abnormal Cell Kind » qui regroupe les éléments d'anatomo-cyto-pathologie, « Anatomy Kind » qui regroupe les concepts d'anatomie ou « Clinical or Research Activity Kind » pour les concepts relatifs à la recherche ou à la clinique. Associés à ces concepts, on retrouve des propriétés comme le terme préféré (preferred term), des synonymes, des définitions, et d'autres informations complémentaires.

En plus de la relation hiérarchique de type *is\_a*, le NCIt possède plus de 90 types de relations entre les concepts, comme par exemple *Disease\_Has\_Abnormal\_Cell* pour qualifier une relation entre une pathologie et un type de cellules anormales ou encore *Disease\_Has\_Primary\_Anatomic\_Site* pour indiquer le site anatomique d'atteinte primitif d'un cancer. Aujourd'hui, le NCIt décrit plus de 200 000 relations entre ses concepts.

Le NCIt comporte à la fois des concepts « pré-coordonnés » (par exemple, *Breast Carcinoma* qui est codé C4872) et des concepts simples pouvant être combinés selon le principe de post-coordination (par exemple, *Breast* qui est codé C12971 peut être combiné à *Carcinoma* qui est codé C2916). L'autre élément majeur du NCIt est la possibilité d'utiliser les relations hiérarchiques et transversales pour associer les concepts entre eux, qu'ils soient dans un même sous-domaine ou pas. Par exemple, le concept *Breast Carcinoma* de l'axe « Disease, Disorder or Finding » possède notamment une relation de type *Disease\_Has\_Primary\_Anatomic\_Site* avec *Breast* de l'axe « Body Part, Organ, or Organ Component ».

Associé au fait que le NCIt est une terminologie disponible gratuitement en « open-data » et continuellement remise à jour, la structure du NCIt en fait une terminologie de choix pour

élaborer une terminologie cible ayant pour but l'utilisation conjointe de données issues de différentes sources en cancérologie. Cependant, le NCIt n'a pas été élaboré pour répondre à cet objectif, il n'est pas exhaustif et des concepts à représenter peuvent manquer. De plus, des erreurs rencontrées lors de son utilisation pratique ont déjà été décrites dans la littérature [23-25], soulignant la nécessité de retravailler cette terminologie, de l'enrichir et de la corriger au fil du travail de mapping, pour la faire évoluer vers la terminologie cible souhaitée (Figure 2).



**Figure 2 : Adaptation du NCIt comme outil permettant l'utilisation conjointe de données de cancérologie issues de différentes sources, initialement codées dans différentes terminologies.**

## 2.4. Choix de la CIM-O3 comme première terminologie à mapper

### 2.4.1. CIM-O3 : une terminologie diagnostique de référence

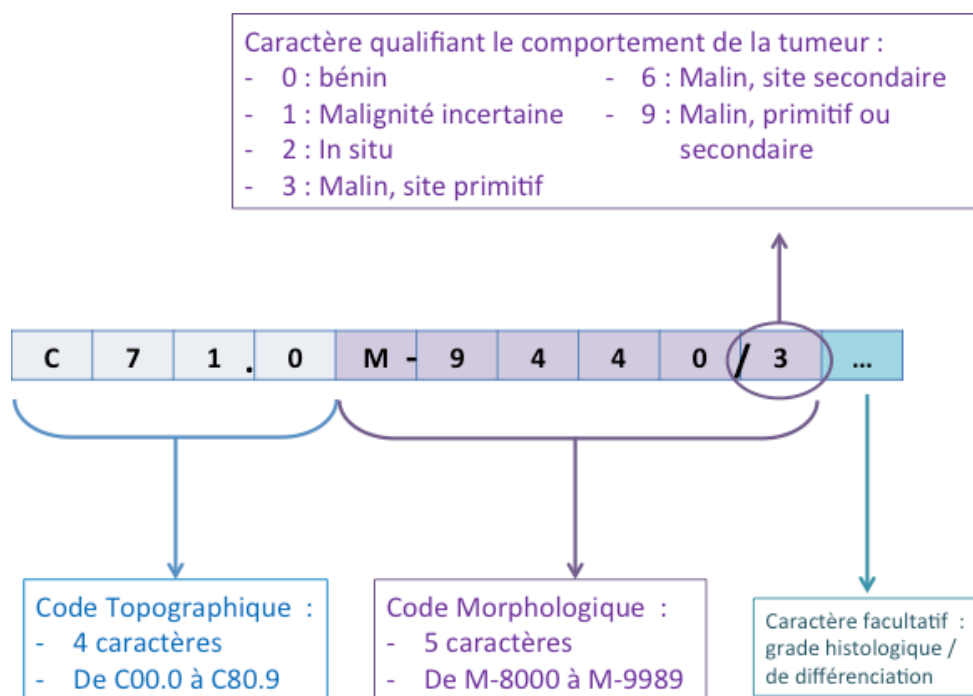
L'une des terminologies de référence utilisée en pratique en cancérologie pour coder le diagnostic de tumeur est la CIM-O3 [19]. La CIM-O3 a été créée par l'OMS et est principalement utilisée par les registres des cancers internationaux et lors des réunions de concertation pluridisciplinaires français pour le codage des diagnostics de tumeurs. Dans le cadre de l'élaboration d'une terminologie cible basée sur le NCIt permettant l'utilisation conjointe de données en cancérologie, le mapping de la CIM-O3 vers le NCIt s'est avéré indispensable.

### 2.4.2. Structure de la CIM-03

La CIM-03 [19] possède deux axes contenant plus de 1500 concepts. Un axe morphologique (1091 concepts), construit à partir de la SNOMED version 3.5 [20], concerne la morphologie de la tumeur et un axe topographique (409 concepts) précise une localisation, le site primitif de la tumeur, construit à partir de la section C de la CIM-10 [16] et complété par l'adjonction des sites de tumeurs hématopoïétiques et réticulo-endothéliales, non présents dans la CIM-10.

Les codes topographiques sont composés de quatre caractères et sont compris entre C00.0 et C80.9. Un point (.) sépare les sous-localisations des catégories codées par les trois premiers caractères. Par exemple, le code C71.0 représente l'hémisphère (.0) cérébral (C71), le code C50.2 représente le quadrant supéro-interne (.2) du sein (C50), le code C34.1 représente le lobe supérieur (.1) du poumon (C34).

Les codes morphologiques comportent cinq chiffres compris entre M-8000/0 et M-9989/3. Les quatre premiers chiffres représentent l'histologie précise. Le cinquième chiffre, derrière la barre oblique (/), est le code de comportement qui indique si la tumeur est maligne, bénigne, in situ ou de caractère malin ou bénin non précisé. Par exemple, le code M-9440/3 représente le glioblastome (M-9440) qui a un caractère « Malin, site primitif » (précisé par le /3). Un caractère supplémentaire (optionnel) indique le grade histo-pathologique ou le degré de différenciation. Pour les lymphomes et les leucémies, cet élément permet d'identifier le type de cellules (T, B, NK ou nulle). En pratique, ce caractère n'est pas utilisé pour les autres types de tumeur.



**Figure 3 : Structure du codage d'un diagnostic par l'intermédiaire de la CIM-03 (exemple du glioblastome de caractère « Malin, site primitif » de l'hémisphère cérébral)**

Le codage d'une tumeur précise dans un organe précis en CIM-O3 correspond à l'agrégation des 4 caractères du code topographique et des 5 caractères du code morphologique, éventuellement suivi d'un caractère facultatif. Par exemple, une tumeur codée C71.0 M-9440/3 correspond à un glioblastome qui a un caractère « Malin, site primitif » de l'hémisphère (.0) cérébral (Figure 3). On peut ainsi caractériser la CIM-O3 comme une terminologie nécessitant une post-coordination pour le codage des tumeurs car le diagnostic complet n'est pas codé tel que dans la terminologie.

L'une des particularités de la CIM-O3 est l'indépendance totale entre ses deux axes : il est théoriquement possible de composer chacun des codes topographiques avec chacun des codes morphologiques, même s'il existe pour certaines morphologies des recommandations sur les topographies qui peuvent la concerner, comme le précise la règle H de codage de la CIM-O3 présentée en Annexe 1. La prise en compte de cette particularité est l'un des enjeux de l'élaboration d'un mapping de qualité de la CIM-O3 vers le NCIt.

## 2.5. Mise à l'épreuve du mapping

### 2.5.1. Confrontation à des données réelles

L'objectif à terme qu'est l'élaboration d'une terminologie cible permettant d'utiliser conjointement les données issues de différentes sources implique de pouvoir représenter prioritairement les données utilisées en pratique.

Ainsi, il est apparu indispensable de confronter le mapping de la CIM-O3 vers le NCIt à des données réelles, afin d'évaluer la possibilité de représenter les diagnostics issues de la clinique, codés en CIM-O3, avec des concepts du NCIt.

La CIM-O3 étant la nomenclature de référence des registres des cancers dans le monde et les données recueillies par un registre étant un reflet exhaustif de l'épidémiologie des cancers du territoire le concernant, l'utilisation des données diagnostiques issus d'un registre des cancers a été choisi pour cette évaluation.

### 2.5.2. Registre général des cancers de la Gironde

Suite au premier Plan Cancer 2003-2007, l'Institut de Santé Publique, d'Épidémiologie et de Développement (ISPED) a été sollicité par l'Institut de Veille Sanitaire (InVS) afin de mettre en place un registre général des cancers en Gironde (RGCG), en renforcement des registres régionaux et nationaux existants, fédérés par le réseau Francim. Après une période de mise en place, le recueil a débuté en janvier 2005 et s'est progressivement intensifié avec une

multiplication des sources d'information. Le RGCG a obtenu sa première qualification par le Comité National des Registres (CNR) fin 2006, puis une deuxième qualification fin 2009.

Le RGCG est situé à l'ISPED, composante de l'Université de Bordeaux, et rattaché au Centre de Recherche INSERM U 897. Il est placé sous la direction du Pr Roger SALAMON et la direction adjointe du Dr Gaëlle COUREAU. L'équipe réunit également le Pr Simone MATHOULIN-PELISSIER, consultante pour la recherche, un chef de projets, M. Brice AMADEO, un informaticien, des Attachées de Recherche Clinique (ARC), deux secrétaires et un interne de santé publique.

L'objectif principal du RGCG est d'estimer l'incidence des cancers en Gironde et son évolution dans le temps, grâce à la collecte exhaustive de tous les cas de cancers incidents dans le département. Cet objectif entre dans le cadre de la surveillance épidémiologique des cancers et les données du registre contribuent à alimenter la base de données nationale. En terme de recherche, les données issues du registre représentent une ressource précieuse. Le registre participe ainsi aux études coordonnées par le réseau Francim et par d'autres acteurs de la recherche, comme l'INSERM ou les départements de recherche universitaires. La Gironde est le plus grand département français, comptant une population d'environ 1,5 millions d'habitants (données au 1<sup>er</sup> janvier 2014, Insee). Par ailleurs, la répartition de la population selon l'âge est analogue à celle de l'ensemble de la population française (données au 1<sup>er</sup> janvier 2014, Insee).

#### 2.5.2.1. Sources d'information du RGCG

Le recueil des données est assuré par le médecin coordonnateur et les ARC du RGCG, en collaboration avec les équipes des différents registres spécialisés existant dans le département. Il s'agit d'un recueil actif de l'ensemble des nouveaux cas auprès des laboratoires d'anatomie pathologique, grâce à un passage régulier du personnel du registre, complété par le croisement de la base du registre avec des bases de données issues de différents organismes comme les établissements de santé par l'intermédiaire du PMSI ou par l'accès aux dossiers patients, les caisses d'assurance maladie, les bases nationales type Répertoire National d'Identification des Personnes Physiques (RNIPP), le Réseau de Cancérologie d'Aquitaine, etc.

#### 2.5.2.2. Critères d'inclusion des cas

Après le croisement des données collectées auprès des sources, le RGCG inclut dans sa base de données :

- Toute nouvelle tumeur :
  - maligne invasive, quelle que soit la localisation
  - in situ ou superficielle de la vessie
  - *borderline* des ovaires

- in situ des seins
- Diagnostiquée à compter du 01/01/2005
- Chez un sujet résidant en Gironde

Sont exclus du registre :

- Les carcinomes baso-cellulaires
- Les récives tumorales
- Les tumeurs de patients ayant refusé leur inclusion au sein de la base du registre
- Les tumeurs enregistrées par les registres spécialisés du département : tumeurs du Système Nerveux Central (SNC), hémopathies malignes et tumeurs d'une personne mineure au moment du diagnostic

Les registres n'incluent que des cas de tumeurs primitives.

Les recommandations au sujet de ces critères d'inclusion sont élaborées par le réseau Francim à partir des recommandations de l'*International Agency for Research on Cancer* (IARC), dans le but d'unifier le recueil des registres du réseau [26].

#### 2.5.2.3. Codage du diagnostic

L'ensemble des informations relatives à une tumeur, recueillies de façon automatique ou manuelle, codées ou textuelles, permet aux ARC de coder le diagnostic de la tumeur en CIM-O3, comme le recommande l'IARC [26]. Le codage se fait par synthèse des informations par un ARC, spécialisé dans le codage des tumeurs d'un ou deux organes. Ainsi, les tumeurs sont codées par un expert de la localisation de la tumeur, ce qui permet une grande efficacité du codage.

Pour chaque tumeur sont enregistrés :

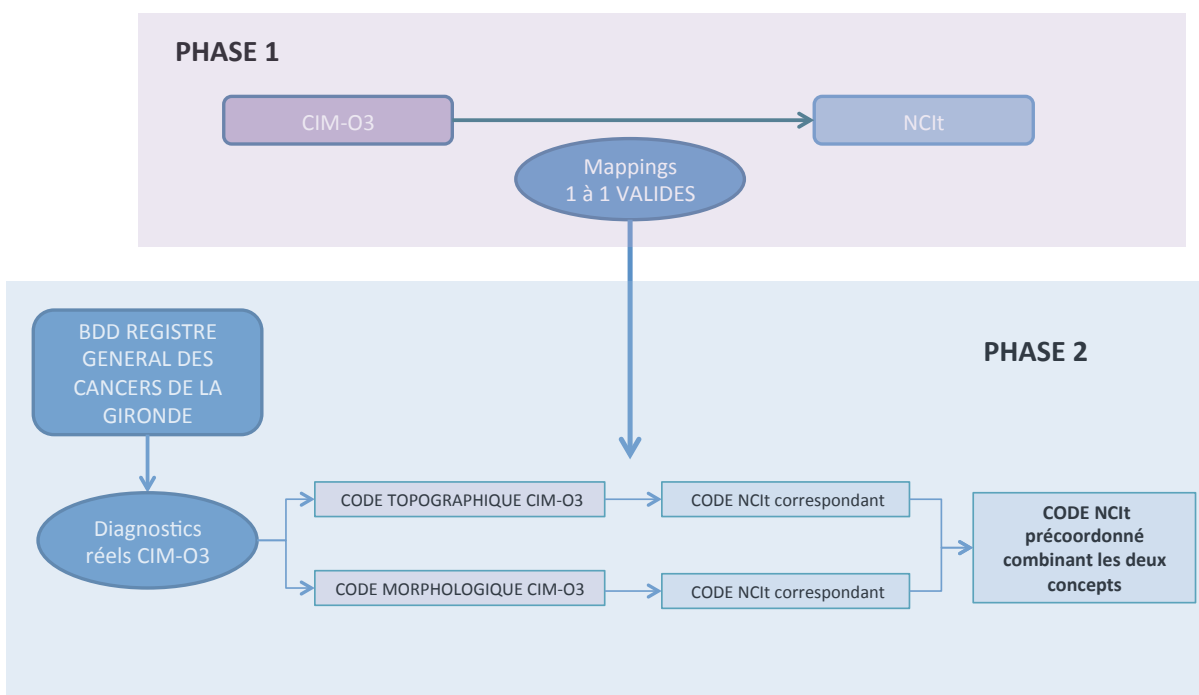
- Un code CIM-O3 représentant la morphologie de la tumeur
- Un code CIM-O3 représentant la topographie de la tumeur

## 2.6. Objectif

L'objectif principal de ce travail était d'évaluer la possibilité de représenter les données diagnostiques d'un registre général des cancers de l'adulte (hors hémopathies et tumeurs du SNC) sous forme de termes pré-coordonnés du NCIt, grâce à un mapping validé établi entre la CIM-O3 et le NCIt et en utilisant les relations existant entre les concepts du NCIt.

### 3. MATERIEL ET METHODE

Pour la réalisation de ce travail, une première phase a consisté en l'élaboration d'un mapping de la CIM-O3 vers le NCIt, validé par une équipe de médecins. La seconde phase s'est attachée à évaluer la possibilité de représenter les diagnostics réels du RGCG, codés en CIM-O3, avec des concepts du NCIt. Pour cela, les termes du NCIt de topographie et de morphologie ont d'abord été recherchés à partir des termes CIM-O3 utilisés par le RGCG en utilisant le mapping obtenu lors de la première phase. Ensuite, ces termes du NCIt ont été combinés afin de voir si le terme pré-coordonné correspondant au diagnostic CIM-O3 précis complet existait dans le NCIt. La *Figure 4* présente une vue générale du travail et, en particulier, l'articulation des phases l'une par rapport à l'autre.



**Figure 4: Schéma général regroupant la phase 1 de mapping de la CIM-O3 vers le NCIt et son exploitation dans la phase 2 pour associer les diagnostics CIM-O3 du RGCG à des termes pré-coordonnés du NCIt**

#### 3.1. Mapping de la CIM-O3 vers le NCI thesaurus

Pour respecter l'objectif de ce travail, il faut obtenir une correspondance unique et juste de chaque concept source vers un concept cible. En d'autres termes, il faut que chaque code source pointe vers un seul et unique code de la terminologie cible, et que les deux concepts mis en relation soient exactement équivalents. On cherche à établir un mapping dit de type « 1 à 1 » pour chaque concept car des mappings multiples, de type « 1 à n », où il existe plusieurs

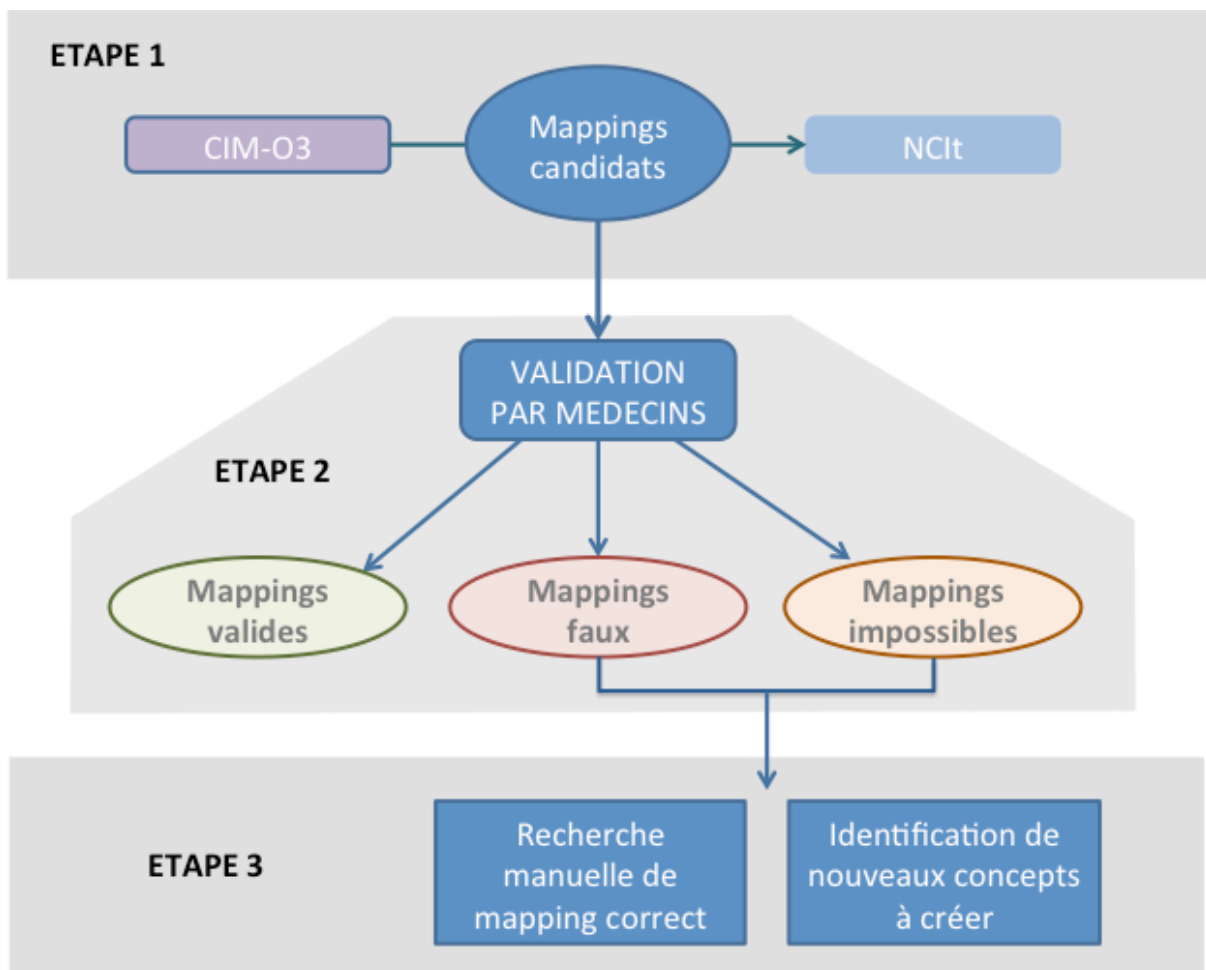


codes cibles pour un même code initial, ne sont pas exploitables automatiquement : ils nécessitent une intervention manuelle pour choisir le bon code cible.

Ainsi, la première phase de ce travail cherchait à élaborer un mapping de la CIM-O3 vers le NCIt avec un mapping de type « 1 à 1 » pour chaque code CIM-O3 vers un unique code NCIt.

Le travail de mapping s'est fait en trois étapes (Figure 5) :

- recherche de mappings candidats de façon automatisée ;
- analyses automatique puis manuelle des mappings candidats par une équipe de médecins ;
- identification des erreurs éventuelles à corriger et des concepts à créer dans la terminologie cible.



*Figure 5 : Elaboration du mapping 1 à 1 valide de la CIM-O3 vers le NCIt en trois étapes.*

Ce travail a été réalisé à partir des versions anglaises des deux terminologies, afin de pouvoir comparer plus aisément les termes entre eux, sans avoir à effectuer une traduction des termes, le NCIt n'étant disponible qu'en anglais.

Les versions des terminologies utilisées étaient :

- La version 2013\_06D du NCIt ;
- La troisième version (2011) de la CIM-O3.

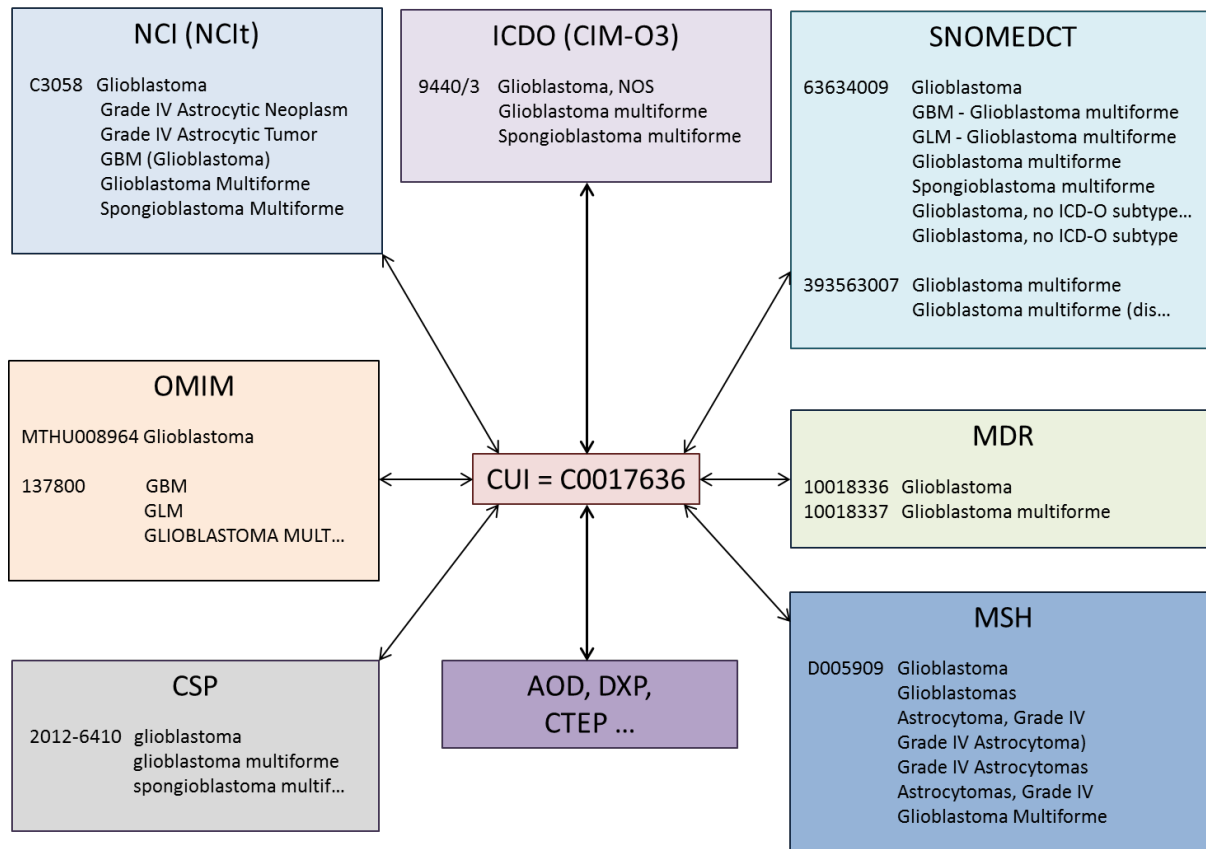
### 3.1.1. Etape 1 : recherche de mappings candidats

La recherche de mapping candidats de façon automatique a été réalisée par l'intermédiaire de deux outils du NCI métathésaurus [27].

#### 3.1.1.1. NCI metathesaurus

Le NCI metathesaurus est un système utilisé dans le spectre de la cancérologie regroupant plus de 75 terminologies différentes, en version anglaise. Il a été élaboré par le *National Cancer Institute* américain à partir de l'UMLS metathesaurus [13] enrichi par des terminologies biomédicales spécifiques à la cancérologie, comme la CIM-O3. On retrouve environ quatre millions de termes représentant deux millions de concepts biomédicaux dans le NCI metathesaurus. Ces concepts sont reliés entre eux par plus de vingt-deux millions de relations, présentes dans les différentes terminologies. Les termes et codes issus des terminologies sources représentant un même concept sont identifiés dans le NCI metathesaurus grâce à un code unique : le Concept Unique Identifier [28] ou CUI (*Figure 6*).

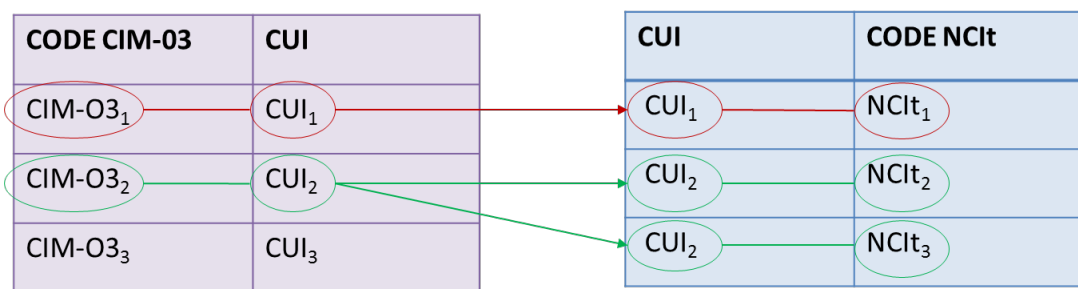
Pour chaque concept de chaque terminologie, donc pour chaque code de chaque terminologie, présent dans le NCI metathesaurus, il existe un certain nombre de termes représentatifs du concept, dont l'un d'entre eux est désigné comme étant le terme préféré. Chaque terme présent dans le NCI metathesaurus est normalisé, c'est-à-dire qu'il est traité par un algorithme de normalisation textuelle [29] permettant de supprimer les variations d'écriture d'un même mot (suppression des accents, de la ponctuation, des majuscules, de certains mots vides comme « the », abolition de l'ordre des mots, etc.). Le résultat de cette normalisation est le Language Unique Identifier [28] ou LUI, regroupant sous un même code l'ensemble des variations lexicales d'un terme. Par exemple, les termes Bladder neck, Bladder, neck, Neck of bladder ou Neck of the Bladder seront normalisés de la même façon et auront le même LUI (L1010165).



**Figure 6 : Exemple de regroupement des termes et codes de différentes terminologies au sein d'un même CUI dans le NCI metathesaurus.**

### 3.1.1.2. Mappings candidats

Pour établir des mappings candidats entre la CIM-O3 et le NCIt, le CUI et le LUI ont été utilisés. Pour l'approche par le CUI, il s'agissait de rechercher des codes CIM-O3 et NCIt ayant le même CUI (Figure 7). L'approche par le LUI était semblable : il s'agissait de chercher les codes NCIt pour lesquels au moins l'un des termes NCIt avait le même LUI qu'un des termes du code CIM-O3.



**Figure 7 : mappings candidats obtenus à partir du CUI du NCI metathesaurus**

A l'issue de cette recherche, ont été obtenus des codes CIM-O3 pour lesquels nous n'avions aucun mapping candidat (mappings de type « 1 à 0 »), des codes CIM-O3 pour lesquels nous avons un seul mapping candidat (mappings de type « 1 à 1 ») et des codes CIM-O3 pour lesquels il y avait plusieurs mappings candidats (mappings de type « 1 à n »).

### 3.1.2. Etape 2 : évaluation des mappings candidats

L'ensemble des mappings candidats retrouvés grâce aux CUI et aux LUI ont ensuite été analysés par plusieurs médecins de santé publique assistés par des anatomo-pathologistes si nécessaire, suivant des règles de validation préétablies. La validation se faisait au regard des termes préférés des concepts.

Un mapping a ainsi été jugé valide s'il associait :

- deux concepts dont les termes préférés étaient des synonymes médicaux exacts, en accord avec le dictionnaire de référence *Concise medical dictionary* [30].
- un concept CIM-O3 dont le terme préféré portait l'acronyme NOS (*not otherwise specified* ou en français SAI, sans autre indication) et un concept NCIt dont le terme préféré est exactement équivalent, même si l'acronyme NOS n'apparaissait pas.

Un mapping a été jugé invalide s'il associait :

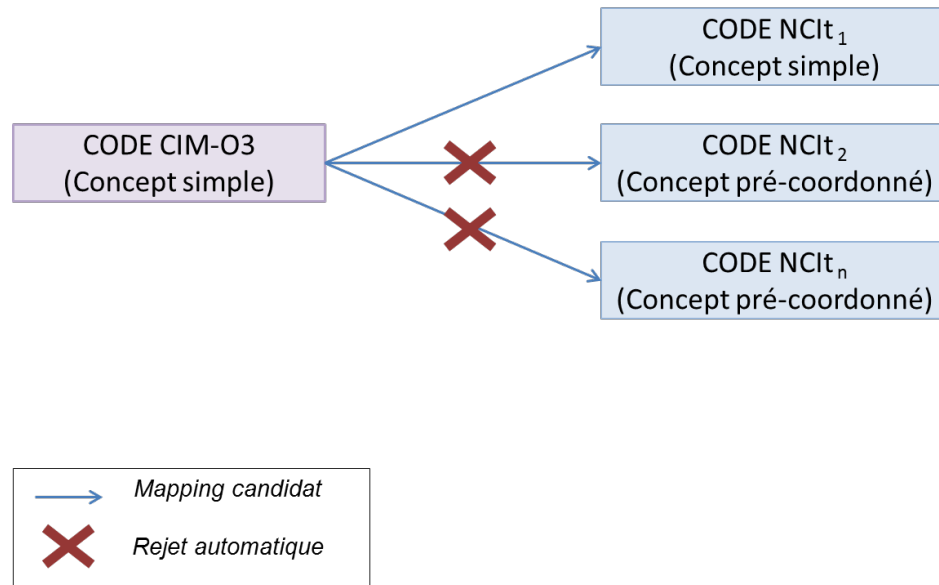
- un concept CIM-O3 (concept simple de pathologie ou de morphologie) et un concept NCIt pré-coordonné (concept d'une pathologie précise dans un organe donné, dont la liste a été générée selon la méthode décrite en annexe 2).
- un concept CIM-O3 et un concept NCIt qui s'avérait être un parent du concept CIM-O3, et non un concept exactement équivalent.

En accord avec ces règles, un premier filtrage automatique a été réalisé éliminant les mappings d'un concept CIM-O3 vers un concept NCIt pré-coordonné (*Figure 8*).

Les mappings candidats restants ont été évalués manuellement par deux médecins différents en aveugle, chacun validant ou rejetant chaque mapping. Dans le cas d'un désaccord entre ces deux validations ou dans le cas où les médecins n'arrivaient pas à trancher (par exemple pour des concepts extrêmement spécifiques comme ceux des tumeurs hématopoïétiques), un médecin spécialiste en anatomo-pathologie a été consulté.

Pour les mappings candidats où plusieurs codes NCIt étaient possibles pour un même code CIM-O3, un consensus pour déterminer le meilleur mapping a été recherché.

A l'issue de cette étape, des mappings « 1 à 1 » valides ont été obtenus, mais aussi des codes pour lesquels aucun mapping n'a été trouvé et enfin des codes pour lesquels les mappings retrouvés ont été jugés faux.



**Figure 8 : Processus de filtrage automatique des mappings entre un concept CIM-O3 et un concept NCIt pré-coordonné**

### 3.1.3. Etape 3 : enrichissement manuel de la terminologie cible

Pour les codes non mappés à l'issue de la validation par l'équipe de médecins, une recherche manuelle de mappings potentiels a été réalisée. Les codes ne pouvant être mappés du fait de l'absence de concepts correspondants dans le NCIt ont été identifiés dans l'optique d'un enrichissement ultérieur de la terminologie cible.

## 3.2. Représentations des diagnostics de tumeurs du RGCG au sein du NCIt

En utilisant le mapping validé obtenu lors de la première phase et les relations décrites dans le NCIt, nous avons cherché à associer les diagnostics CIM-O3 effectivement utilisés par RGCG à des diagnostics NCIt pré-coordonnés.

### 3.2.1. Extraction des diagnostics de cancers codés dans la base du registre

Pour chaque tumeur présente dans le RGCG, un diagnostic unique de la tumeur est codé en CIM-O3, grâce à l'association d'un code topographique CIM-O3 et d'un code morphologique CIM-O3.

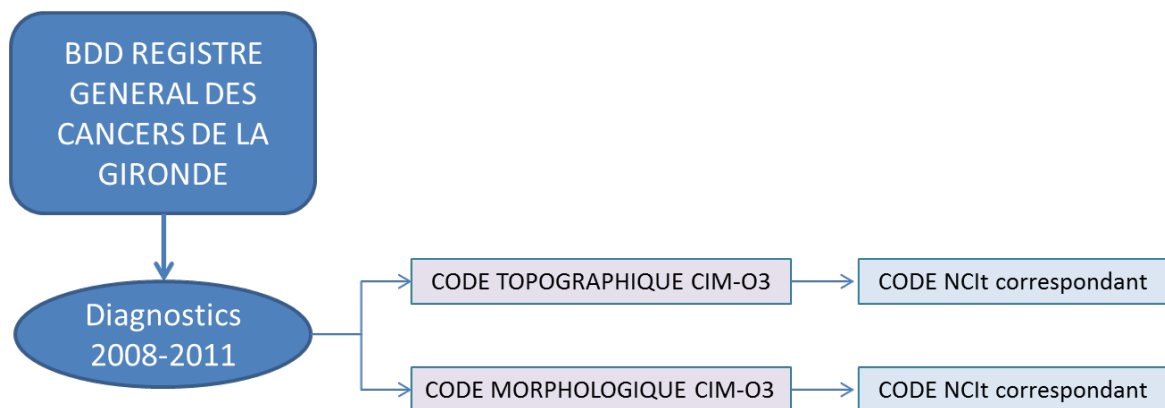
L'ensemble des diagnostics, soit les associations entre un code topographique CIM-O3 et un code morphologique CIM-O3, de chaque tumeur incluse ayant une date d'incidence comprise entre le 01/01/2008 et le 31/12/2012, ont été extraits de la base de données du RGCG. La

borne basse de la période d'inclusion a été choisie car, avant le 01/01/2008, la transmission d'information relative aux cas de cancers par les sources du registre n'était pas exhaustive. Pour la borne haute, la fin de la dernière année où les cas incidents ont été entièrement traités et validés par le registre a été choisie.

Aucune donnée nominative n'a été extraite de la base de données du RGCG, l'extraction ne concernait que les diagnostics des tumeurs ainsi qu'un identifiant aléatoire de tumeur.

### 3.2.2. Mapping des topographies et morphologies CIM-O3 vers le NCIt

Pour chaque diagnostic, une correspondance avec un code NCIt a été recherchée pour le code topographique CIM-O3 et pour le code morphologique CIM-O3, à l'aide des mappings validés de la CIM-O3 vers le NCIt obtenu lors de la première phase de ce travail (Figure 9).



**Figure 9 : Mapping des diagnostics présents dans la base de données du registre général des cancers de la Gironde vers le NCIt.**

A l'issue de ce mapping, une distinction a été faite entre :

- Les diagnostics pour lesquels le code topographique trouvait une correspondance dans le NCIt ;
- Les diagnostics pour lesquels le code morphologique trouvait une correspondance dans le NCIt ;
- Les diagnostics pour lesquels les deux codes, morphologique et topographique, trouvaient une correspondance dans le NCIt, appelés par la suite « Mapping Complet ».

### 3.2.3. Recherche de concepts pré-coordonnés

Pour l'ensemble des diagnostics issus du RGCG classés en « Mapping complet », le ou les concepts NCIt pré-coordonnés correspondants ont été recherchés dans une liste extraite du NCIt suivant la méthode décrite en Annexe 2. Cette liste met en correspondance les diagnostics pré-coordonnés et les concepts simples de topographie et morphologie qui les composent.

Les résultats de cette recherche ont été examinés par une équipe de médecins en aveugle pour validation puis comparés. En cas de désaccord, un anatomo-pathologiste pouvait être consulté afin d'obtenir un consensus sur la validité du mapping.

### 3.3. Technologies utilisées

L'ensemble des données ont été stockées dans une base de données MySQL, sur un serveur sécurisé de l'Equipe de Recherche en Informatique Appliquée à la Santé (ERIAS) du centre INSERM U897. La base de données a été interrogée par l'intermédiaire de phpMyAdmin® et de programmes développés en PHP®. La version OWL (*Web Ontology Language*) du NCIt a été visualisée par l'intermédiaire du logiciel Protege® 4.0.2.

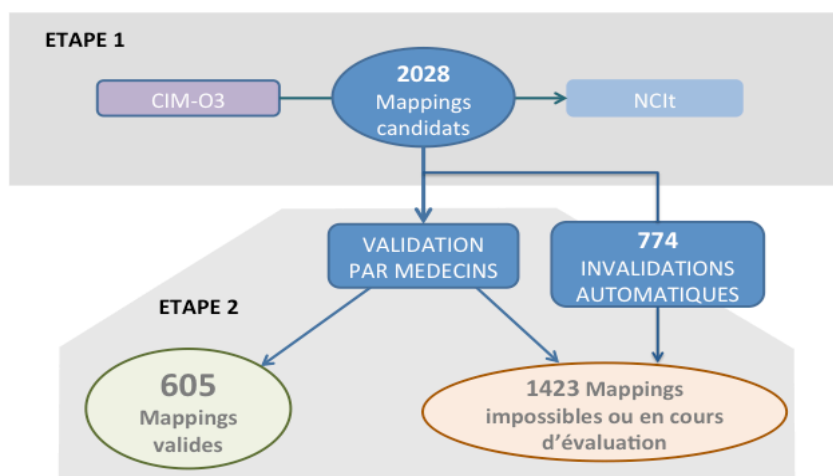
## 4. RESULTATS

### 4.1. Mapping de la CIM-O3 vers le NCI thesaurus

La recherche de correspondances de la CIM-O3 dans le NCIt par l'intermédiaire des CUI et des LUI a résulté en 2028 mappings candidats, dont 754 pour des codes CIM-O3 topographiques et 1274 pour des codes CIM-O3 morphologiques. Près de 7% des concepts CIM-O3 ont été mappés uniquement par l'intermédiaire du CUI, 3% uniquement par l'intermédiaire du LUI et 90% par les deux méthodes. Par exemple, le concept de naevus cellulaire ballon (Balloon cell nevus - code CIM-O3 8722/0) n'a pas pu être mappé à un concept NCIt par l'intermédiaire du CUI mais retrouve une correspondance avec le concept Balloon Nevus Cell (code NCIt C36863) par l'intermédiaire du LUI.

L'invalidation automatique des mappings entre un concept CIM-O3 et un concept NCIt pré-coordonné a permis d'écarter 774 mappings (38,2% des mappings candidats).

Au 31 janvier 2015, la validation par l'équipe de médecins avait permis d'obtenir 605 mappings de type « 1 à 1 » valides entre un code CIM-O3 et un code NCIt (soit 40,3 % des 1500 codes CIM-O3). Plus précisément, ces mappings correspondaient à 342 (31,3%) des 1091 codes CIM-O3 morphologiques et 263 (64,3%) des 409 codes CIM-O3 topographiques. La quasi-totalité des 895 codes CIM-O3 restants sont des concepts qui n'existent pas tels que dans le NCIt et constituent donc des concepts qu'il serait nécessaire de créer. Par ailleurs, certains concepts étaient des concepts obsolètes. Quelques mappings candidats étaient encore en cours d'évaluation par une anatomo-pathologiste car très spécialisés. Par exemple, la relation entre les concepts Solid pseudopapillary tumor (8452/1) et Papillary Cystic Neoplasm (C4179) est difficile à juger et nécessite l'avis d'un anatomo-pathologiste spécialisé. Concernant la validation par l'équipe de médecins, une variabilité inter-évaluateurs de l'ordre de 10% a été observée avant la recherche de consensus et la consultation du spécialiste en anatomo-pathologie. Les résultats du mapping de la CIM-O3 vers le NCIt sont résumés dans la *Figure 10*.



*Figure 10 : Résultats du mapping de la CIM-O3 vers le NCIt*



## 4.2. Représentation des diagnostics de tumeurs du RGCG au sein du NCIt

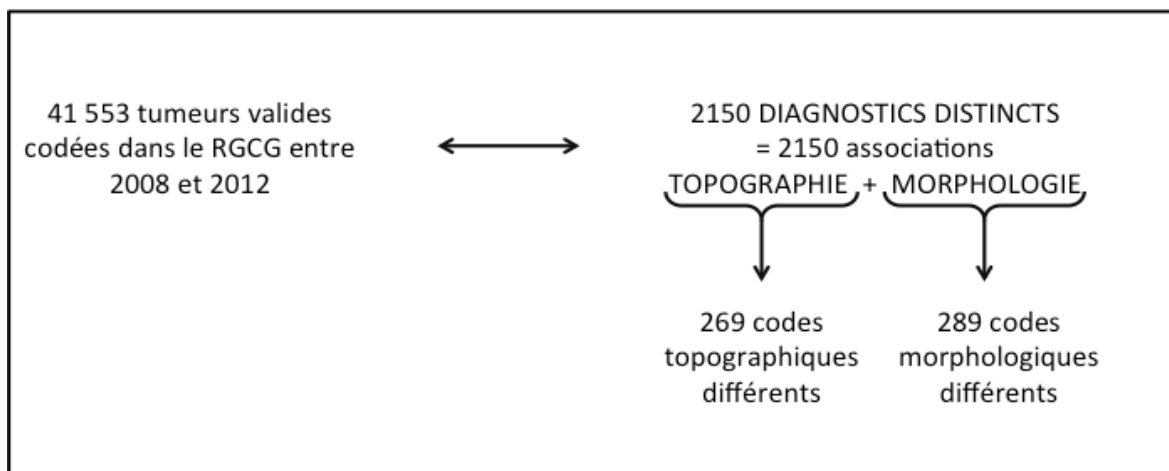
### 4.2.1. Description des diagnostics de tumeurs codés dans le RGCG

Au 20 janvier 2015, date de l'extraction des diagnostics de tumeurs de la base du RGCG, 41 553 tumeurs ayant une date de diagnostic entre 2008 et 2012 et étant codées en CIM-O3 ont été extraites.

L'ensemble de ces tumeurs était codé à l'aide d'un jeu de 2150 associations distinctes entre un code topographique CIM-O3 et un code morphologique CIM-O3. En d'autres termes, le registre a utilisé 2150 diagnostics distincts pour coder l'ensemble des tumeurs incluses durant la période analysée.

Ces 2150 diagnostics correspondaient à des combinaisons entre (*Figure 11*) :

- 269 codes topographiques distincts ;
- 289 codes morphologiques distincts.



**Figure 11 : Analyse et décomposition des diagnostics de tumeurs utilisés par le Registre Général des Cancers de la Gironde (RGCG) entre 2008 et 2012**

En considérant l'ensemble des tumeurs du RGCG durant la période d'inclusion, 5160 des 41 553 tumeurs (12,4%) étaient codées comme étant des Adénocarcinome, SAI (8140/3) de la Prostate (C61.9), 2098 des 41 553 tumeurs (5%) étaient codées comme étant des Adénocarcinome canalaire infiltrant, SAI (8500/3) du Quadrant supéro-externe du sein (C50.4) et 1222 des 41 553 tumeurs (2,9%) étaient codées comme étant des Adénocarcinome canalaire infiltrant, SAI (8500/3) du Quadrant supéro-interne du sein (C50.2).

Les topographies les plus représentées étaient la Prostate (5246 tumeurs - 12,9%) le Quadrant supéro-externe du sein (2907 tumeurs - 7%) et le Lobe supérieur du poumon, SAI (2374 tumeurs - 5,7%). Les morphologies les plus représentées étaient l'Adénocarcinome, SAI (8699 tumeurs - 20,9%), l'Adénocarcinome canalaire infiltrant (5447 tumeurs - 13,1%) et le Carcinome épidermoïde ou spinocellulaire (3627 tumeurs - 8,7%).

## 4.2.2. Mapping des diagnostics CIM-O3 vers des concepts NCIt simples

### 4.2.2.1. Résultat du mapping

La recherche de codes NCIt simples (i.e., non pré-coordonnés) correspondant soit à un code CIM-O3 topographique, soit à un code morphologique a permis de retrouver sur les 2150 diagnostics distincts utilisés :

- un code topographique NCIt pour 1664 diagnostics (77,4%). Ceci correspond à 193 codes NCIt distincts retrouvés pour 269 codes CIM-O3 utilisés par le RGCG (71,7%)
- un code morphologique NCIt pour 1563 diagnostics (72,7%). Ceci correspond à 138 codes NCIt distincts retrouvés pour 289 codes CIM-O3 utilisés par le RGCG (47,8%).

En considérant l'ensemble des tumeurs du RGCG durant la période d'inclusion, on retrouvait ainsi :

- Une correspondance NCIt pour le code topographique de 35 057 des 41 553 tumeurs, soit 84,4 % des tumeurs pour lesquelles on retrouvait un équivalent NCIt au code topographique ;
- Une correspondance NCIt pour le code morphologique de 27 274 des 41 553 tumeurs, soit 65,6 % des tumeurs pour lesquelles on retrouvait un équivalent NCIt au code morphologique.

Les trois topographies les plus représentées dans le registre pour la période d'inclusion trouvaient toutes une correspondance dans le NCIt. En revanche, seules deux des trois morphologies les plus représentées trouvaient une correspondance dans le NCIt (Table 1).

**Table 1 : Résultat du mapping de la CIM-O3 vers le NCIt des topographies et morphologies les plus représentées dans le registre général des cancers de la Gironde pour la période 2008-2012**

|             | CODE CIM-O3 | Libellé CIM-O3                                   | Nombre de diagnostics concernés dans le registre (total = 41553) | Mapping du concept vers le NCIt |
|-------------|-------------|--------------------------------------------------|------------------------------------------------------------------|---------------------------------|
| TOPOGRAPHIE | C61.9       | PROSTATE                                         | 5346                                                             | Oui                             |
|             | C50.4       | QUADRANT SUPÉRO-EXTERNE DU SEIN                  | 2907                                                             | Oui                             |
|             | C34.1       | LOBE SUPÉRIEUR DU POUMON SAI                     | 2374                                                             | Oui                             |
| MORPHOLOGIE | 8140/3      | ADÉNOCARCINOME, SAI                              | 8699                                                             | Oui                             |
|             | 8500/3      | ADÉNOCARCINOME CANALAIRE INFILTRANT, SAI         | 5447                                                             | Non                             |
|             | 8070/3      | CARCINOME ÉPIDERMOÏDE, CARCINOME SPINOCELLULAIRE | 3627                                                             | Oui                             |

Au total, 1195 diagnostics distincts (55,6% des 2150 diagnostics distincts utilisés par le RGCG) ont pu trouver une correspondance complète par association d'un code NCIt correspondant à la topographie et d'un code NCIt correspondant à la morphologie et ont donc été classés comme « Mapping complet ». Ainsi, 21 694 tumeurs (52,2% des 41 553 tumeurs du RGCG) avaient un mapping complet de leur diagnostic.

Il existait 469 diagnostics (21,8% des 2150 diagnostics distincts du RGCG) pour lesquels seul un code topographique a été retrouvé et 368 diagnostics (17,1% des diagnostics du RGCG) pour lesquels seul un code morphologique a été retrouvé.

Au final, il y avait 118 diagnostics (5,5% des 2150 diagnostics distincts du RGCG) auxquels aucun code NCIt n'a pu être associé, correspondant à 952 tumeurs (2,3% des 41 553 tumeurs du RGCG).

#### 4.2.2.2. Description des mappings manquants

##### **Codes topographiques**

Sur les 269 codes topographiques utilisés par le RGCG, 76 codes n'ont pas pu être mappés au NCIt. Parmi ces 76 codes, 12 codes correspondaient à des « sous-codes » de la peau (Code topographique C44) utilisés localement par le RGCG pour préciser la région de la peau correspondante. Ce « sous-codage » était réalisé par adjonction d'un caractère supplémentaire à un code CIM-O3 préexistant (codage à cinq caractères au lieu des quatre caractères de la CIM-O3). Le « sous-code » qui en résultait n'était donc pas un code de la CIM-O3 mais un code propre au registre. Par exemple, la peau du nez était codée C44.32, la peau du front C44.31, au

lieu d'être regroupés sous un même code moins précis de la CIM-O3, le C44.3, qui correspond à la peau d'autres régions de la face. De la même façon, le registre codait C44.62 la peau de la face dorsale de la main et C44.63 la peau de la face palmaire de la main, ou C44.69 la peau du bras, SAI plutôt que de coder l'ensemble de ces éléments sous le code CIM-O3 (très vaste) C44.6 de la peau du membre supérieur et de l'épaule.

Parmi les 64 codes restants non retrouvés dans le NCIt, 32 codes correspondaient à des codes en « .8 » représentant des localisations contiguës d'un organe ou d'une région. Par exemple, le code C50.8 Localisations contiguës du sein n'a pas pu être mappé.

Les 32 autres codes correspondaient à des concepts de granularité fine qui n'existaient pas dans le NCIt ou à des concepts de classifications anatomo-pathologiques non représentées dans le NCIt. Par exemple, le concept Canal biliaire extra-hépatique (Extrahepatic bile duct – C24.0) n'a pas pu être mappé car aucun concept d'une finesse équivalente n'existe dans le NCIt.

### **Codes morphologiques**

En analysant les codes morphologiques ne retrouvant pas de correspondance dans le NCIt, un diagnostic n'ayant pas de morphologie indiquée dans la base du RGCG a été identifié. Il s'agissait d'un diagnostic de tumeur localisée au niveau de la Plèvre.

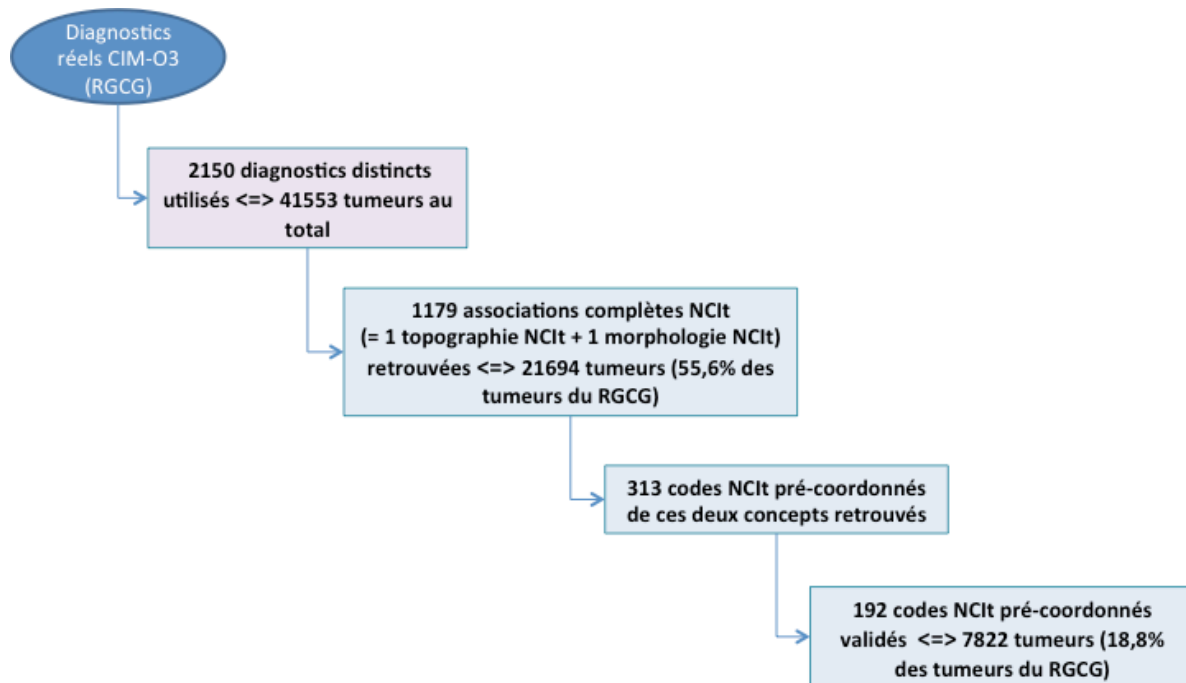
L'analyse des codes morphologiques manquants a montré qu'une très grande majorité d'entre eux correspondaient soit à des concepts de granularité très fine qui n'existaient pas dans le NCIt, soit à des concepts n'existant pas en tant que concept simple dans le NCIt, soit à des concepts de classifications anatomo-pathologiques non représentées dans le NCIt. Par exemple, le concept de Carcinome cribriiforme intra-épithélial (Cribriform carcinoma in situ - 8201/2) n'a pas pu être mappé à un concept NCIt simple : le seul mapping candidat disponible l'est avec un concept pré-coordonné.

### **4.2.3. Diagnostics pré-coordonnés**

Pour les 1195 diagnostics classés en tant que « Mapping complet », la recherche du diagnostic CIM-O3 pré-coordonné par l'intermédiaire des relations décrites dans le NCIt a permis de retrouver 313 codes NCIt pré-coordonnés candidats (*Figure 12*).

Après évaluation par l'équipe de médecins, 192 diagnostics pré-coordonnés ont été validés (61,4% des diagnostics pré-coordonnés candidats, soit 8,9% des diagnostics distincts utilisés par le RGCG) et 38 diagnostics pré-coordonnés potentiels étaient encore en cours d'examen par un anatomo-pathologiste. Les 83 diagnostics pré-coordonnés invalidés l'ont été car ils ne correspondaient pas exactement à l'association des deux concepts simples. Par exemple, les concepts Pancreas (C12393) et Carcinoma (C2916) pointaient vers Pancreatic Carcinoma with Mixed Differentiation (C45843) qui n'est pas la représentation exacte du carcinome du pancréas.

Au total, 7822 tumeurs (18,8% des 41 553 tumeurs du RGCG) avaient un diagnostic qui pointait vers un concept NCIt pré-coordonné valide (Figure 12).



**Figure 12 : Résultats de la recherche de concepts NCIt pré-coordonnés correspondant aux diagnostics CIM-O3 utilisés par le Registre Général des Cancers de la Gironde (RGCG)**

## 5. DISCUSSION

### 5.1. Synthèse des résultats et analyse

#### 5.1.1. Mapping de la CIM-O3 vers le NCIt

Ce travail a permis d'obtenir 605 mappings de type 1 à 1 de la CIM-O3 vers le NCIt, validés par une équipe de médecins. Aucune évaluation de ce type n'avait été réalisée précédemment et ce jeu de données peut aujourd'hui constituer un gold standard pour l'évaluation éventuelle de nouvelles techniques de mapping. Il sera, dans cet objectif, mis à disposition du public.

L'utilisation d'une combinaison entre LUI et CUI comme méthode de recherche de mappings candidats est une technique peu décrite dans la littérature. En effet, si l'usage du CUI est le plus courant pour établir des mappings à partir d'un outil libre d'accès et préexistant [31-33], le LUI est moins connu et souvent négligé au profit de techniques d'alignement lexicales plus complexes [34-35]. Les deux techniques s'avèrent être complémentaires car près de 10% des concepts CIM-O3 ne retrouveraient pas de mapping candidat avec l'utilisation isolée d'une seule de ces deux méthodes (7% des mappings candidats retrouvés uniquement par la méthode basée sur le CUI et 3% uniquement par la méthode exploitant le LUI). Cependant, pour établir des mappings potentiels de la CIM-O3 vers le NCIt, l'utilisation combinée des CUI et LUI génère un grand nombre de mappings non valides si l'objectif est d'obtenir des mappings « 1 à 1 » entre deux termes strictement équivalents : 38,2% des mappings candidats associent un concept simple et un concept pré-coordonné et moins de 30% des mappings candidats ont été validés. Ceci démontre que les outils largement utilisés comme référence pour établir des mappings ne sont pas utilisables tels que si l'objectif est d'obtenir des mappings de type « 1 à 1 » entre des concepts de deux terminologies. De plus, même s'il semblerait au premier abord que les mappings candidats obtenus conjointement par les deux méthodes soient des mappings avec un fort potentiel de validité, la validation manuelle a montré que beaucoup s'avèrent être des mappings entre deux concepts n'étant pas exactement équivalents.

La génération d'une liste de concepts pré-coordonnés selon la méthode de Jouhet (2015) (annexe 2) a permis d'appliquer la règle selon laquelle, dans le cadre d'un mapping entre deux concepts exactement équivalents, il est par définition impossible d'avoir une relation entre un concept simple et un concept pré-coordonné. Cette technique de filtrage des mappings candidats par l'utilisation d'une liste de termes pré-coordonnés pour éliminer des mappings faux est une technique originale qui n'avait jamais été décrite dans la littérature. Elle a permis d'épargner un temps d'examen considérable par l'équipe de médecins. Cependant, il faut noter que la liste des termes pré-coordonnés n'est probablement pas exhaustive car elle est construite uniquement à partir des relations décrites entre les concepts dans le NCIt. Cette technique de filtrage peut donc encore être améliorée.

Près de 900 codes n'ont pas pu être mappés car les concepts CIM-O3 à mapper n'étaient pas représentés dans le NCIt. Ceci met en évidence que, malgré le nombre important de concepts présents dans le NCIt (plus de 43 000), cette terminologie n'ayant pas été développée spécifiquement pour la représentation des diagnostics, elle nécessite un enrichissement.

Un repérage manuel de la position éventuelle dans la hiérarchie du NCIt de ces concepts non mappés et des relations éventuelles de ces concepts avec d'autres concepts du NCIt est en cours de réalisation, avec pour objectif d'enrichir la terminologie cible et d'augmenter le nombre de mappings valides de type « 1 à 1 » entre la CIM-O3 et le NCIt.

## 5.1.2. Représentation des diagnostics du RGCG

### 5.1.2.1. Diagnostics du RGCG

L'extraction des diagnostics de la base du RGCG a montré que le registre a utilisé 269 codes topographiques et 289 codes morphologiques, soit 558 codes de la CIM-O3, pour coder les diagnostics de cancers sur la période étudiée. Ce nombre relativement faible peut s'expliquer par le fait qu'il s'agit d'un registre général des cancers qui, d'une part, n'enregistre que les tumeurs malignes et, d'autre part, n'enregistre pas les cancers spécialisés comme les tumeurs du système nerveux central, les tumeurs de l'enfant, les hémopathies malignes ou les mésothéliomes. Ainsi, le RGCG écarte automatiquement les tumeurs codées à l'aide de 115 codes morphologiques correspondant à des hémopathies malignes, 11 codes morphologiques correspondant aux mésothéliomes, 17 codes correspondant aux carcinomes baso-cellulaires (dont 7 bénins) et 6 codes topographiques correspondant au système nerveux central. Ainsi, 149 codes spécialisés sont écartés automatiquement, auxquels s'ajoutent les codes correspondant aux tumeurs bénignes. De plus, quelques concepts retrouvés dans la CIM-O3 sont des concepts obsolètes qui ont aujourd'hui été remplacés et ne sont plus utilisés lorsqu'il s'agit de coder un diagnostic de tumeur. Il paraît donc logique que le RGCG ne les ait pas utilisés pour coder les tumeurs durant la période d'inclusion. Notons cependant qu'il est important de conserver et mapper les concepts obsolètes dans l'optique d'une utilisation de données anciennes.

### 5.1.2.2. Mapping des composantes topographie et morphologie vers le NCIt

Si l'on considère la première phase de ce travail, qui consistait à retrouver pour chaque élément de diagnostic codé en CIM-O3 une correspondance dans le NCIt, les performances du mapping valide établi en amont étaient faibles car seuls 55,6% des diagnostics distincts trouvaient une correspondance à la fois pour leur code morphologique et pour leur code topographique. Cependant, l'examen des diagnostics ne trouvant pas de correspondance permet d'expliquer en partie ce faible taux et de proposer des solutions simples pour l'améliorer.

Concernant les 12 « sous-codes » topographiques de la peau ne retrouvant pas de correspondance dans le NCIt, deux solutions sont possibles afin d'obtenir des mappings valides :

- L'ensemble des « sous-codes » peuvent être traités sans précision, comme le code parent correspondant, donc comme le code CIM-O3 réel. Par exemple, le « sous-code » C44.63 de la peau de la face palmaire serait traité comme C44.6 (peau du membre supérieur et de l'épaule).
- Si ces « sous-codes » traduisent une nécessité pour les registres de pouvoir coder plus précisément ces topographies, de nouveaux codes devraient être créés dans la terminologie cible (et probablement dans la terminologie source, *i.e.*, la CIM-O3). La création de ces concepts enfants serait aisée car les concepts parents sont identifiés et existent dans la terminologie cible.

Après consultation du RGCG, la pratique d'enregistrement des localisations de la peau sous cinq caractères étant purement locale et n'étant que très peu utilisée en pratique, la première solution a été retenue et on soulignera que les pratiques de codage du registre ont évolué en ce sens.

Lorsque l'on examine les 73 codes topographiques restants ne retrouvant pas de correspondance, il s'agit de concepts n'existant pas tels quels dans le NCIt, et donc de concepts qu'il faudrait créer au sein de la terminologie cible.

Pour les 32 codes de localisation contiguë (codés « .8 »), la création de nouveaux codes de façon automatisée dans la terminologie cible peut sembler facile si l'on considère que le concept doit être un concept frère de la localisation de laquelle il est contigu. Cependant, une localisation contiguë n'étant pas un organe, elle ne peut pas être représenté dans l'axe « organe » (et ne peut donc pas non plus être frère d'un concept organe). Une solution afin d'obtenir des mappings valides serait de créer une nouvelle classe « localisations contiguës » dans la terminologie cible, qui serait un concept parent des concepts correspondant aux concepts CIM-O3 codés en « .8 ».

Concernant les codes morphologiques, la grande majorité des absences de mapping sont dues à des divergences entre les différentes classifications utilisées par les spécialistes. Toutes les classifications ne sont pas représentées dans le NCIt et pour beaucoup de classifications, il est impossible d'établir des liens car les spécialistes eux-mêmes n'appliquent pas les mêmes règles pour classer les tumeurs. Là encore, le manque de consensus international est un frein à l'utilisation conjointe des données.

Si l'on considère l'ensemble des tumeurs du RGCG, les performances du mapping étaient également faibles puisque seules 52,2% des 41 553 tumeurs du RGCG trouvaient un mapping complet de leur diagnostic. Ceci s'explique en partie par le fait que la deuxième morphologie la plus représentée (Adénocarcinome canalaire infiltrant, SAI – codé 8500/3, concernant plus de 13% des tumeurs du registre) ne retrouvait pas de correspondance avec un concept simple



NCIt. En effet, dans le NCIt, ce concept n'existe que sous la forme de pathologies pré-coordonnées. Ainsi, l'utilisation du CUI a résulté en des mappings potentiels qui ont été invalidés automatiquement après utilisation de la règle selon laquelle, dans le cadre d'un mapping entre deux concepts exactement équivalents, il est par définition impossible d'avoir une relation entre un concept simple et un concept pré-coordonné. Ce concept est donc un concept à créer tel que dans la terminologie cible, et à mettre en relation avec l'ensemble des diagnostics pré-coordonnés existant qui s'y rapportent.

Ce dernier élément met en évidence l'une des faiblesses majeures des terminologies diagnostiques prenant le parti de coder les tumeurs de façon pré-coordonnée. En effet, lorsque ce choix est fait, un grand nombre de concepts sont nécessaires à la représentation du panel de tumeurs concernant une morphologie précise, et le risque d'oubli de représentation d'une tumeur est d'autant plus important. De plus, ce choix applique d'importantes contraintes pour la maintenance de la terminologie, par exemple lors d'une évolution des classifications anatomo-pathologiques.

#### 5.1.2.3. Recherche des diagnostics pré-coordonnés

Moins de 9% des diagnostics distincts utilisés par le RGCG ont trouvé une correspondance avec un concept NCIt pré-coordonné. Cependant, si l'on considère l'ensemble des tumeurs du RGCG sur la période d'inclusion, ces diagnostics représentent 18,8% des tumeurs. Si ce chiffre témoigne là encore d'un rendement faible, il est néanmoins le témoin de l'une des vocations du NCIt, qui est d'éditer une terminologie calquée sur la pratique et évoluant en fonction des besoins de ses utilisateurs [36], car les tumeurs présentes en tant que concept pré-coordonné sont les tumeurs les plus fréquemment rencontrées en pratique. Au cours de cette recherche de diagnostics « pré-coordonnés » correspondant aux diagnostics codés par le RGCG, l'une des difficultés a été la sélection d'un seul diagnostic lorsque plusieurs diagnostics « pré-coordonnés » étaient retrouvés. Ainsi, il semble là encore impossible de s'affranchir d'une étape de sélection manuelle.

Une approche « similaire » avait déjà été réalisée par Burgun et Bodenreider [33] qui cherchaient à reconstruire des diagnostics par post-coordination (associant topographie et morphologie) pour les mapper vers le NCIt, par l'intermédiaire du CUI du NCI metathesaurus. Si l'objectif était différent, les causes des inconsistances révélées restent semblables, avec principalement la différence de granularité entre les terminologies et l'amalgame entre morphologie et pathologie. Les apports principaux du présent travail résident dans la confrontation à des données réelles et la validation consensuelle à chaque étape par une équipe de médecins, même si cette dernière entraîne une minimisation des performances des étapes en aval car seules les topographies et morphologies mappées de manière valide sont ensuite pré-coordonnées. De plus, si l'on cherche à obtenir une équivalence exacte entre les

concepts, la validation manuelle rejette près de deux tiers des diagnostics pré-coordonnés candidats, expliquant les écarts entre les performances de ce travail et d'autres tentatives de recherche de mappings pré-coordonnés.

L'ensemble de ce travail a permis de mettre en évidence un certain nombre d'inconsistances au sein même du NCIt, pour la plupart déjà décrites dans la littérature [23-25]. Par exemple, il existe un amalgame entre Adénocarcinome et Carcinome. En effet, certains concepts simples d'Adénocarcinome sont mis en relation avec des concepts pré-coordonnés correspondant à un Carcinome et inversement. Ainsi, le concept de Colon Signet Ring Cell Adenocarcinoma est mis en relation avec le concept morphologique simple de Signet Ring Cell Carcinoma tandis que le concept morphologique simple de Clear Cell Adenocarcinoma est mis en relation avec le concept pré-coordonné de Multilocular Cystic Renal Cell Carcinoma. Par définition, un Adénocarcinome et un Carcinome ne peuvent pas être considérés comme des concepts équivalents car le premier est plus spécifique que le second. Les incohérences identifiées seront communiquées aux équipes du NCIt.

## 5.2. Limites potentielles

### 5.2.1. Variabilité inter-humaine

La validation en aveugle par des médecins a mis en évidence une variabilité inter-évaluateurs de l'ordre de 10%, majoritairement pour ce qui concerne les concepts spécialisés. Cette variabilité est décrite dans la littérature, avec la description de plusieurs méthodes de correction et de prise en compte de cette variabilité [37], notamment la recherche de consensus et la consultation d'un troisième observateur, qui ont été toutes deux mises en œuvre dans ce travail afin de réduire au maximum l'influence du hasard et de la subjectivité dans la validation des mappings. Cependant, il reste impossible de s'affranchir totalement de cette variabilité d'interprétation et certains mappings pourraient être considérés comme faux par d'autres évaluateurs.

### 5.2.2. Choix de validation sur les termes préférés

La validation par les médecins a été réalisée selon la concordance entre les termes préférés. Ce choix était responsable de quelques absences de mappings candidats quand les termes préférés n'étaient pas précisés dans le NCIt. Lorsque plusieurs termes préférés étaient indiqués, les évaluateurs ont choisi le terme le plus représentatif du concept selon eux.

Le choix de validation sur la base des termes préférés semble être la solution la plus consensuelle et la plus valide. Cependant, il a été noté dans certains cas que malgré une correspondance exacte des termes préférés et une équivalence apparente à la lecture des

définitions de concepts du NCIt, il arrivait que la structure même des terminologies implique des divergences des concepts. C'est, par exemple, le cas lorsqu'un concept CIM-O3 et un concept NCIt partagent le même terme préféré et une définition semblable, mais lorsque le NCIt donne comme synonyme du terme préféré un enfant du concept. On retrouve également ces « fausses équivalences » car la CIM-O3 distingue formellement les tumeurs secondaires (codées grâce au comportement « /6 ») des autres tumeurs, obligatoirement primitives, ce qui n'est pas le cas du NCIt : on peut ainsi sous le même terme préféré et une définition semblable avoir un concept CIM-O3 de tumeur primitive exclusivement et un concept NCIt de tumeur sans précision du caractère primitif ou secondaire.

### 5.2.3. Utilisation de diagnostics issus d'un registre général

Les diagnostics utilisés pour la deuxième phase du travail étaient issus d'un registre général des cancers. Ainsi, dans le jeu de données initial, aucun diagnostic de tumeur de ce type n'était présent. Ces diagnostics étant des diagnostics souvent très spécialisés, il est possible que les performances de leur mapping soient plus faibles que celles des diagnostics plus communs.

## 5.3. Perspectives

L'objectif à terme de ce projet est la production d'un outil répondant aux besoins de consensus dans la représentation des données diagnostiques, exprimés par les acteurs du domaine de la cancérologie. En effet, l'abandon des multiples terminologies utilisées dans ce domaine au profit d'une terminologie commune consensuelle ne paraît pas réalisable, d'une part, parce qu'il semble impossible d'obtenir ce consensus et, d'autre part, parce que cela nécessiterait de recoder l'ensemble des données anciennes pour garantir leur exploitation future. La solution d'une terminologie pivot mettant en relation ces concepts semble donc à ce jour la meilleure réponse à cette problématique.

Si la masse colossale de travail de mapping pour mettre en correspondance chaque concept de chaque terminologie diagnostique en cancérologie paraît utopique, la confrontation à des données réelles issues d'un registre a permis de mettre en évidence les priorités de mapping à établir, en fonction des besoins des praticiens et chercheurs du domaine.

L'expérience de ce travail a également permis de révéler le réel intérêt des terminologies diagnostiques construites sur deux axes (topographie et morphologie), à partir desquels il est plus aisé de construire de façon automatisée les diagnostics complets.

## 6. CONCLUSION

Ce travail a permis d'obtenir un jeu de données composé de 605 mappings de type « 1 à 1 » de la CIM-O3 vers le NCIt, validés par une équipe de médecins, qui pourra servir de *gold standard* pour des études futures. La mise à l'épreuve de ce jeu de données sur des diagnostics réels issus du RGCG a montré la nécessité de compléter ce mapping et d'étoffer le NCIt afin d'élaborer une terminologie pivot fiable et complète, permettant l'utilisation conjointe de données diagnostiques codées dans des terminologies différentes. Cependant, les analyses des résultats ont permis de mettre en évidence le besoin d'améliorer cette terminologie pivot et de proposer des pistes dans ce sens. De plus, les techniques développées lors de la réalisation de ce travail pourront être utilisées ultérieurement pour l'élaboration de mappings vers d'autres terminologies, ou pour la correction de mappings existants, permettant un enrichissement supplémentaire de la terminologie pivot.

## BIBLIOGRAPHIE

1. Hedden H. Controlled vocabularies, thesauri, and taxonomies. *The Indexer*. 1 mars 2008;26(1):33-4.
2. L'Homme M-C. *La terminologie: principes et techniques*. PUM; 2004. 283 p.
3. Shortliffe EH, Cimino JJ. *Biomedical informatics: computer applications in health care and biomedicine*. New York, NY: Springer; 2006.
4. Jj C. Review paper: coding systems in health care. *Methods Inf Med*. déc 1996;35(4-5):273-84
5. Levy B. Evolving to clinical terminology. *J Healthc Inf Manag*. 2004;18(3):37-43.
6. Kalra D, Fernando B, Morrison Z, Sheikh A. A review of the empirical evidence of the value of structuring and coding of clinical information within electronic health records for direct patient care. *Informatics in Primary Care*. 1 mai 2012;20(3):171-80.
7. Rose JS, Fisch BJ, Hogan WR, Levy B, Marshal P, Thomas DR, et al. Common medical terminology comes of age, Part One: Standard language improves healthcare quality. *J Healthc Inf Manag*. 2001;15(3):307-18.
8. Dudeck J. Changes in Medical Documentation over the Last Five Decades. *Stud Health Technol Inform*. 2008;134:255-63.
9. Rodeghiero F, Stasi R, Gernsheimer T, Michel M, Provan D, Arnold DM, et al. Standardization of terminology, definitions and outcome criteria in immune thrombocytopenic purpura of adults and children: report from an international working group. *Blood*. 12 mars 2009;113(11):2386-93.
10. Blair DR, Wang K, Nestorov S, Evans JA, Rzhetsky A. Quantifying the Impact and Extent of Undocumented Biomedical Synonymy. *PLoS Comput Biol*. 25 sept 2014;10(9):e1003799.
11. Cimino JJ. Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. *Methods Inf Med*. nov 1998;37(4-5):394-403.
12. Ingenerf J, Reiner J, Seik B. Standardized terminological services enabling semantic interoperability between distributed and heterogeneous systems. *International Journal of Medical Informatics*. 1 déc 2001;64(2):223-40.
13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucl Acids Res*. 1 janv 2004;32(suppl 1):D267-70.
14. Zhang S, Bodenreider O. Experience in Aligning Anatomical Ontologies. *Int J Semant Web Inf Syst*. 2007;3(2):1-26.

15. Association pour le Développement de l'Informatique en Cytologie et Anatomopathologie (ADICAP). Thesaurus de la Codification ADICAP. Édition 1999-Version° 4-99/1.pdf. Dernière mise à jour : février 1999 (disponible sur <http://www.adicap.asso.fr/>).
16. International Classification of Diseases (ICD). <http://www.who.int/classifications/icd/en/> (consulté le 01/12/2014)
17. NCI Metathesaurus. <http://ncimeta.nci.nih.gov/ncimbrowser/> (consulté le 20/02/2015)
18. Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thesaurus and Ontology. Web Semantics: Science, Services and Agents on the World Wide Web [Internet] (consulté le 20/02/2015).  
Disponible sur : <http://www.websemanticsjournal.org/index.php/ps/article/view/27>
19. Fritz AG. International Classification of Diseases for Oncology: ICD-O. World Health Organization; 2000. 252 p.
20. Côté RA, Robboy S. Progress in Medical Information Management: Systematized Nomenclature of Medicine (SNOMED). JAMA. 1980;243(8):756-762.
21. Hole WT, Srinivasan S. Discovering missed synonymy in a large concept-oriented Metathesaurus. Proc AMIA Symp. 2000;354-8.
22. Miller U, Teitelbaum R. Pre-coordination and post-coordination: Past and future. Knowledge organization. 2002;29(2):87-93.
23. Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods Inf Med. déc 2004;44(4):498-507.
24. Fragoso G, de Coronado S, Haber M, Hartel F, Wright L. Overview and utilization of the NCI Thesaurus. Comp Funct Genom. 1 déc 2004;5(8):648-54.
25. Kumar A, Smith B. Oncology Ontology in the NCI Thesaurus. In: Miksch S, Hunter J, Keravnou ET, éditeurs. Artificial Intelligence in Medicine [Internet]. Springer Berlin Heidelberg; 2005 [cité 24 févr 2015]. p. 213 - 20. Disponible sur: [http://link.springer.com/chapter/10.1007/11527770\\_30](http://link.springer.com/chapter/10.1007/11527770_30)
26. Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG, ed. Cancer registration: principles and methods. Vol. 95. IARC, 1991.
27. NCI Metathesaurus. <http://ncimeta.nci.nih.gov/ncimbrowser/> (consulté le 25/02/2015)
28. Unique Identifiers in the Metathesaurus [Internet]. Disponible sur: [http://www.nlm.nih.gov/research/umls/new\\_users/online\\_learning/Meta\\_005.html](http://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_005.html) (consulté le 25/02/2015)
29. Lexical Variant Generation (LVG) [Internet]. Disponible sur: [http://www.nlm.nih.gov/research/umls/new\\_users/online\\_learning/LEX\\_004.htm](http://www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_004.htm) (consulté le 25/02/2015)

30. Martin EA. Concise medical dictionary [Internet]. Oxford; New York: Oxford University Press; 2010 [cited 2015 Aug 5]. Available from: [http://www.oxfordreference.com/views/BOOK\\_SEARCH.html?book=t60](http://www.oxfordreference.com/views/BOOK_SEARCH.html?book=t60)
31. Fung KW, Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. AMIA AnnuSymp Proc. 2005;266-70.
32. Sarkar IN, Cantor MN, Gelman R, Hartel F, Lussier YA. Linking Biomedical Language Information and Knowledge Resources: GO and UMLS. Pac Symp Biocomput. 2003;439-50.
33. Burgun A, Bodenreider O. Issues in integrating epidemiology and research information in oncology: experience with ICD-O3 and the NCI Thesaurus. AMIA AnnuSymp Proc. 2007;85-9.
34. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;17-21.
35. Wang, Yefeng, et al. "Linguistic mapping of Terminologies to SNOMED CT." Semantic Mining Conference on SNOMED CT. 2006.
36. de Coronado S, Wright LW, Fragoso G, Haber MW, Hahn-Dantona EA, Hartel FW, et al. The NCI Thesaurus quality assurance life cycle. J Biomed Inform. 2009 Jun;42(3):530–9.
37. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice. SAGE Publications; 2014. 833 p.

## LISTE DES ANNEXES

|                                                                                                                                                          |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| ANNEXE 1 : REGLE H de la CIM-O3.....                                                                                                                     | 41 |
| ANNEXE 2 : Evaluation of the National Cancer Institute thesaurus for integrating disease terminologies in oncology (article en cours de soumission)..... | 43 |



## LISTE DES FIGURES

|                                                                                                                                                                                                               |    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figure 1 : Hiérarchie de concepts au sein d'une terminologie et granularité. ....                                                                                                                             | 4  |
| Figure 2 : Adaptation du NCIt comme outil permettant l'utilisation conjointe de données de cancérologie issues de différentes sources, initialement codées dans différentes terminologies. ....               | 9  |
| Figure 3 : Structure du codage d'un diagnostic par l'intermédiaire de la CIM-O3 (exemple du glioblastome de caractère « Malin, site primitif » de l'hémisphère cérébral) .....                                | 10 |
| Figure 4: Schéma général regroupant la phase 1 de mapping de la CIM-O3 vers le NCIt et son exploitation dans la phase 2 pour associer les diagnostics CIM-O3 du RGCG à des termes pré-coordonnés du NCIt..... | 14 |
| Figure 5 : Elaboration du mapping 1 à 1 valide de la CIM-O3 vers le NCIt en trois étapes. ....                                                                                                                | 15 |
| Figure 6 : Exemple de regroupement des termes et codes de différentes terminologies au sein d'un même CUI dans le NCI metathesaurus.....                                                                      | 17 |
| Figure 7 : mappings candidats obtenus à partir du CUI du NCI metathesaurus.....                                                                                                                               | 17 |
| Figure 8 : Processus de filtrage automatique des mappings entre un concept CIM-O3 et un concept NCIt pré-coordonné .....                                                                                      | 19 |
| Figure 9 : Mapping des diagnostics présents dans la base de données du registre général des cancers de la Gironde vers le NCIt.....                                                                           | 20 |
| Figure 10 : Résultats du mapping de la CIM-O3 vers le NCIt .....                                                                                                                                              | 22 |
| Figure 11 : Analyse et décomposition des diagnostics de tumeurs utilisés par le Registre Général des Cancers de la Gironde (RGCG) entre 2008 et 2012 .....                                                    | 23 |
| Figure 12 : Résultats de la recherche de concepts NCIt pré-coordonnés correspondant aux diagnostics CIM-O3 utilisés par le Registre Général des Cancers de la Gironde (RGCG).....                             | 27 |

## LISTE DES TABLES

|                                                                                                                                                                                                   |    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Table 1 : Résultat du mapping de la CIM-O3 vers le NCIt des topographies et morphologies les plus représentées dans le registre général des cancers de la Gironde pour la période 2008-2012 ..... | 25 |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|

## ABREVIATIONS

|             |                                                                                                                    |
|-------------|--------------------------------------------------------------------------------------------------------------------|
| ADICAP      | Thésaurus créée par l'Association pour le Développement de l'Informatique en Cytologie et en Anatomie Pathologique |
| ARC         | Attaché de Recherche Clinique                                                                                      |
| CIM-9 / -10 | Classification Internationale des Maladies                                                                         |
| CIM-O3      | Classification internationale des maladies pour l'oncologie                                                        |
| CNR         | Comité National des Registres                                                                                      |
| CUI         | Concept Unique Identifier                                                                                          |
| ERIAS       | Equipe de Recherche en Informatique Appliquée à la Santé                                                           |
| Francim     | Réseau France Cancer Incidence et Mortalité                                                                        |
| IARC        | International Agency for Research on Cancer                                                                        |
| Insee       | Institut national de la statistique et des études économiques                                                      |
| INSERM      | Institut national de la santé et de la recherche médicale                                                          |
| InVS        | Institut de Veille Sanitaire                                                                                       |
| ISPED       | Institut de Santé Publique, d'Épidémiologie et de Développement                                                    |
| LUI         | Language Unique Identifier                                                                                         |
| NCIt        | NCI Thesaurus                                                                                                      |
| OMS         | Organisation Mondiale de la Santé                                                                                  |
| PMSI        | Programme de Médicalisation du Système d'Information                                                               |
| RGCG        | Registre général des cancers en Gironde                                                                            |
| RNIPP       | Répertoire National d'Identification des Personnes Physiques                                                       |
| SAI         | Sans autre indication (équivalent français de NOS, not otherwise specified)                                        |

## CONVENTION D'ECRITURE

Les concepts des terminologies sont notés en police *American Typewriter*.

## ANNEXE 1 : REGLE H de la CIM-O3.

**RÈGLE H. Utiliser le code topographique fourni lorsque la localisation topographique n'est pas indiquée par le diagnostic. Ce code topographique ne doit pas être utilisé lorsqu'un autre site d'origine est connu pour la tumeur.**

Certains termes désignant des tumeurs indiquent implicitement un site ou tissu d'origine particulier. Le tableau 23 en présente quelques exemples.

| <b>Tableau 23. Exemples de termes morphologiques avec localisation associée</b> |                                          |                                                   |           |                                        |                  |
|---------------------------------------------------------------------------------|------------------------------------------|---------------------------------------------------|-----------|----------------------------------------|------------------|
| <b>Code</b>                                                                     | <b>Terme</b>                             | <b>Topographie CIM-O</b>                          |           | <b>Autres</b>                          |                  |
| <b><u>Morphologique</u></b>                                                     | <b><u>Terme</u></b>                      | <b><u>(localisation primitive habituelle)</u></b> |           | <b><u>localisations primitives</u></b> |                  |
| M-9510/3                                                                        | Rétinoblastome                           | C69.2                                             | Rétine    | -                                      |                  |
| M-8170/3                                                                        | Carcinome hépatocellulaire               | C22.0                                             | Foie      | -                                      |                  |
| M-8090/3                                                                        | Carcinome basocellulaire                 | C44.-                                             | Peau      | C51.-                                  | Vulve            |
|                                                                                 |                                          |                                                   |           | C60.-                                  | Pénis            |
|                                                                                 |                                          |                                                   |           | C63.2                                  | Scrotum          |
|                                                                                 |                                          |                                                   |           | C61.9                                  | Prostate         |
| M-9530/0                                                                        | Méningiome                               | C70.-                                             | Méninges  | -                                      |                  |
| M-938--M-948                                                                    | Gliomes                                  | C71.-                                             | Encéphale | C72.0                                  | Moelle épinière  |
| M-8500/3                                                                        | Adénocarcinome canalaire infiltrant, SAI | C50.-                                             | Sein      | C07.9                                  | Glande parotide  |
|                                                                                 |                                          |                                                   |           | C08.-                                  | Glande salivaire |
|                                                                                 |                                          |                                                   |           | C25.-                                  | Pancréas         |
|                                                                                 |                                          |                                                   |           | C61.9                                  | Prostate         |
| M-8470/3                                                                        | Cystadénocarcinome mucineux, SAI         | C56.9                                             | Ovaire    | C25.-                                  | Pancréas         |
|                                                                                 |                                          |                                                   |           | C34.-                                  | Poumon           |

Afin de faciliter le codage de ces termes, le code topographique est, le cas échéant, ajouté entre parenthèses aussi bien dans la liste numérique des codes morphologiques que dans l'index alphabétique. Il peut arriver aussi que le code topographique soit proposé après un terme d'en-tête codé par trois chiffres, auquel cas il s'applique à tous les termes répertoriés sous ce premier.

À “ carcinome basocellulaire ” (tableau 23), le code topographique correspondant à la peau (C44.-) est fourni, à l'exception du quatrième caractère qui reste à déterminer. Le trait ( \_ ) qui suit le point indique qu'il existe des sous-localisations pouvant être codées à cet endroit par un quatrième caractère. Pour connaître, en vue du codage, les codes correspondant à des sous-localisations particulières, il convient de se reporter à la liste numérique ou à l'index alphabétique. Un carcinome basocellulaire du visage, par exemple, se code C44.3 (peau du visage), tandis que le même carcinome affectant le bras se code C44.6 (peau du bras). De même, l'emplacement du quatrième caractère du code topographique attribué à “ méningiome ” (C70.-) est laissé vacant dans la mesure où la localisation impliquée peut être les “ méninges cérébrales ” (C70.0), les “ méninges rachidiennes ” (C70.1) ou les “ méninges, SAI ” (C70.9).

**Le code topographique fourni avec le terme morphologique peut être utilisé lorsque la localisation n'est pas indiquée par le diagnostic.** Il existe un grand nombre de termes morphologiques auxquels aucun code topographique n'est attribué parce qu'il s'agit de tumeurs survenant fréquemment dans plus d'un organe ou plus d'une localisation. Pour “ adénocarcinome, SAI ”, par exemple, aucun code

topographique n'est fourni parce que cette tumeur peut avoir pour localisation primitive de nombreux organes différents. Il peut arriver que la localisation indiquée par le diagnostic diffère de celle fournie avec le terme. Le carcinome basocellulaire, par exemple, peut avoir d'autres localisations que la peau à son apparition. **Pour le codage, il convient d'ignorer le code topographique proposé par la CIM-O au profit de celui correspondant à la localisation primitive indiquée par le diagnostic si ceux-ci diffèrent.** Par exemple, le code topographique C50.- (sein) est proposé pour le terme morphologique "carcinome canalaire infiltrant" car il s'agit d'un type de carcinome survenant généralement au niveau du sein. Cependant, si ce terme est utilisé pour désigner un carcinome primitif survenu au niveau du pancréas, le code proposé pour le sein devra être remplacé par le code approprié, à savoir C25.9 (pancréas, SAI). Il faut bien se souvenir que les codes topographiques proposés en association avec les termes morphologiques désignent les localisations d'origine *habituelles* de certaines tumeurs particulières. Plus inhabituel, quoi que possible, est un diagnostic décrivant, par exemple, un "ostéosarcome du rein". Le code topographique correspondant au rein (C64.9) est alors utilisé à la place de "os, SAI" (C41.9) après vérification complète du dossier afin de s'assurer qu'il ne s'agit pas d'un cancer de l'os métastaté au niveau du rein. Une métastase dans le rein d'un cancer de l'os (ostéosarcome) se code C41.9 (os), M-9180/3 (ostéosarcome).

**Termes morphologiques donnant de pseudo-indications topographiques** Certaines tumeurs portent des noms qui semblent fournir des indications sur leur localisation mais le code topographique ne doit pas forcément être déterminé en fonction de cette localisation. Par exemple, le "carcinome des canaux biliaires" (M-8160/3) constitue un type histopathologique précis, fréquemment rencontré au niveau des voies biliaires intra-hépatiques (C22.1) et des voies biliaires extra-hépatiques (C24.0), et ne doit donc pas systématiquement recevoir le code C24.0. Les tumeurs des glandes salivaires accessoires peuvent être observées dans n'importe quelle zone de la cavité buccale et des organes voisins et peuvent être de types histopathologiques multiples, comme "carcinome adénoïde kystique", "tumeur mixte maligne" ou "adénocarcinome, SAI". C'est pour cette raison qu'il n'existe pas de code morphologique distinct propre aux "carcinomes des glandes salivaires accessoires". Dans la mesure où tous les types d'adénocarcinome de la bouche ou de la cavité buccale sont considérés comme ayant pour origine les glandes salivaires accessoires, les mots "glandes salivaires accessoires" doivent être ignorés lorsqu'ils sont rencontrés dans un diagnostic du type "carcinome adénoïde kystique des glandes salivaires accessoires du palais dur". Dans ce cas, le code topographique correspondant à "palais dur" (C05.0) doit être attribué au "carcinome adénoïde kystique" (M-8200/3). Si le diagnostic n'indique aucune localisation d'origine, comme par exemple "adénocarcinome des glandes salivaires accessoires", il convient d'utiliser, pour le codage, le code topographique correspondant à la cavité buccale (C06.9) qui recouvre notamment les "glandes salivaires accessoires, SAI".

## **ANNEXE 2 : Evaluation of the National Cancer Institute thesaurus for integrating disease terminologies in oncology (article en cours de soumission)**

### **Evaluation of the National Cancer Institute thesaurus for integrating disease terminologies in oncology**

Vianney Jouhet, MD<sup>a,b</sup>, Fleur Mougin, PHD<sup>a</sup>, Bérénice Brechat, MD<sup>a,b</sup>, Frantz Thiessard, MD-PHD<sup>a,b</sup>

<sup>a</sup> Equipe de Recherche en Informatique Appliquée à la Santé (ERIAS), INSERM U897, Université de Bordeaux, France

<sup>b</sup> Service d'Information Médicale, Pôle de Santé Publique, CHU de Bordeaux, France

#### **Introduction**

With the scalability of electronic health records (EHRs), the amount of data produced in the medical field is rapidly increasing. These data provide new perspectives to: create and disseminate new knowledge; consider the implementation of personalized medicine; offer to patients the opportunity to be involved in the management of their own medical data [1]. Indeed, secondary use of biomedical data produced throughout patient care remains an essential issue [2] and is the subject of numerous studies since several years [1–6]. From 2007, the American Medical Informatics Association emphasized the value of secondary use of medical data: “*Secondary use of health data can enhance healthcare experiences for individuals, expand knowledge about disease and appropriate treatments, strengthen understanding about the effectiveness and efficiency of our healthcare systems, support public health and security goals, and aid businesses in meeting the needs of their customers*” [4].

In the oncology field, it is necessary to identify and describe incident cancer cases within a population in order to facilitate research and public health monitoring. For instance, cancer registries have to exhaustively record incident cases of cancer in a given territory and this task remains time consuming if manual. As early as 1998, a technical report was drawn up by the International Agency for Research on Cancer describing the methods used by different registries for establishing automated procedures to identify new cases using available data [7]. Methods have been proposed for automatically identifying and registering cancers using structured data indexed with standard terminologies [8–12].

However, multiple actors with many different medical specialties are providing information in EHRs. As a result, within EHRs, data describing diseases are recorded according to multiple heterogeneous terminologies even for a single disease happening to a single patient. For instance, in France, reimbursement data use the 10<sup>th</sup> revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) [13] to describe diseases, whereas pathology data use either ADICAP (a French pathology terminology) or the 3<sup>rd</sup> edition of the International Classification of Diseases for Oncology (ICD-O-3) [14] and data from oncology multidisciplinary meetings use

ICD-O-3. Enabling an integrated access to these disease terminologies may improve automated cancer identification.

Although ICD-10 and ICD-O-3 both describe cancer diseases, they exhibit differences in terms of structure and granularity:

- Structural heterogeneity. ICD-10 is a terminology where each code describes a cancer disease taking into account its topography and morphology (or at least its behavior). Conversely, ICD-O-3 describes diseases by combining two axes, *i.e.*, topography and morphology, which can be coded independently.
- Granular heterogeneity. ICD-O-3 provides a much more specific level (finer grained) than ICD-10 does for describing cancer morphology. As a result, many concepts which can be built by combining a topography and a morphology from ICD-O-3 are not available within ICD-10.

Because a disease cannot be equivalent to only a tumor topography or a tumor morphology, it is obviously not possible to find equivalences between concepts represented within these two terminologies. Thus, it is necessary to identify or to build a resource that would enable the integration of cancer disease terminologies taking into account this structural heterogeneity. To achieve this goal, composite relations must be defined between the involved concepts, such as “*a neoplasm is a disease and has a specified morphology, as well as a specified topography*”. The National Cancer Institute thesaurus (NCIt) “*provides reference terminology covering vocabulary for clinical care, translational and basic research, and public information activities*”<sup>1</sup>. It is described as “*a controlled terminology which exhibits ontology-like properties in its construction and use*” [15]. These characteristics “*open up the possibility [...] in linking together heterogeneous resources created by institutions external to the NCI*” [16]. Thus, the NCIt could be used as a resource to bridge the gap between disease terminologies, which are structurally heterogeneous.

However, since 2005, it has been shown in many occasion that the NCIt remains flawed [16–18] and especially logic-based reasoning over the NCIt should be used cautiously. In [18], Schultz *et al.* discussed that the OWL Description Logic (OWL-DL) version of the NCIt may lead to unexpected results which were not visible due to the lack of use case needing logic-based reasoning over the OWL-DL version of the NCIt. We believe that the integration of heterogeneous terminologies describing neoplastic disease corresponds to such a use case. In order to assess the feasibility of using the NCIt for integrating cancer disease terminologies, we have proposed an evaluation of the ability of the NCIt to link the three axes that refer to diseases as described in ICD-10 and ICD-O-3, *i.e.*, the disease itself as well as its morphology and topography.

---

<sup>1</sup> from <http://ncit.nci.nih.gov/>, visited 2015-01-22

## Background

### ICD-O-3

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 morphology is coded with five digits. The first four digits represent the histological description and the fifth digit indicates the behavior (whether benign or malignant) of a neoplasm. “*The topography code indicates the site of origin of a neoplasm; in other words, where the tumor arose*” [14]. From the ICD-O-3 “point of view”, any morphology code can be associated with any topography code. Some tumor morphologies have a “*usual primary site*” but it is expressly stated that these associations are provided only to help coders and should not be considered as systematic (and unique) topography-morphology combinations. An example is given in [14]: “*An unusual, but possible, example would be the diagnoses “osteosarcoma of kidney”, for which the kidney topography code (C64.9) would be used instead of “bone, NOS” (C41.9)[...]*”. Thus, ICD-O-3 describes a disease by combining the morphology of the tumor and the topography from where the tumor arose. The resulting diseases are not described as such within this terminology.

### ICD-10

The ICD-10 describes each disease as a whole concept entailed by a unique code. For instance, *C50.2: Malignant neoplasm upper-inner quadrant of breast* describes two characteristics of the cancer disease:

- The behavior (*Malignant*) which is part of the morphology description.
- The site of origin (*upper-inner quadrant of breast*) which corresponds to the topography.

### Concepts involved in ICD-10 and ICD-O-3

Within ICD-10 and ICD-O-3, three different kinds of concepts are thus involved:

- The morphology of the tumor, which corresponds to the histological description of the tumor. These concepts are available within the ICD-O-3 morphology axis.
- The topography of the tumor, which corresponds to the site of origin of the tumor. These concepts are available within the ICD-O-3 topography axis.
- The disease, which corresponds to the tumor itself and entails information about both the topography and the morphology of the tumor. These concepts are available as such within ICD-10 and can be built by combining ICD-O-3 topographies and morphologies.

### NCIt

The “*NCI Thesaurus (NCIt)* is NCI’s reference terminology. NCIt provides the concepts used in caCORE and caBIG to establish data semantics. It covers terminology for clinical care, translational and basic research, and public

information and administrative activities. NCIt is also a widely recognized standard for biomedical coding and reference, used by a broad variety of public and private partners both nationally and internationally” [19].

An OWL-DL representation of the NCIt is freely available in Web ontology Language format on the NCI website<sup>2</sup>. Although this OWL-DL representation enable logic-based reasoning, some inconsistencies have been discussed and it has been shown that the NCIt should be used cautiously for this purpose [16–18].

In the NCIt, topographies are described in the *Anatomic structure, system, or substance* axis. Morphologies and diseases are represented within the same hierarchy, subsumed by *Neoplasm*. No axis specific to tumor morphologies is defined. Diseases are modeled as anatomic specializations of morphologies. Thus, diseases are formally described as subclasses of tumor morphologies. For example, *Breast adenocarcinoma is\_a Adenocarcinoma* is stated in:

$$\text{Breast adenocarcinoma} \equiv \text{Adenocarcinoma} \sqcap \text{Breast carcinoma}$$

Some of the NCIt concepts are annotated as being mapped to some ICD-O-3 morphologies. These annotations concern 1,212 concepts mapped to 908 distinct ICD-O-3 morphology codes. For example, *Invasive ductal carcinoma, not otherwise specified* is annotated as being mapped to two ICD-O-3 morphology codes (8500/3 *Infiltrating duct carcinoma, NOS* and 8521/3 *Infiltrating ductular carcinoma*).

## Methods

All the analyses were processed over the OWL-DL representation version of the NCIt (14.11d) available at [http://evs.nci.nih.gov/fp1/NCI\\_Thesaurus/](http://evs.nci.nih.gov/fp1/NCI_Thesaurus/). All DL-queries were executed using the OWL-API v3.4.8 after reasoning with Fact++ v1.6.2.

### Defining a pattern for linking ICD-O-3 and ICD-10

As the main goal is to enable cancer identification, we focused our study on primary tumor descriptions avoiding metastases. In order to link ICD-O-3 and ICD-10 concepts, it is necessary to determine which relationships are involved and how these relationships associate concepts. A particular topography-morphology combination in ICD-O-3 leads to a particular disease description. ICD-O-3 axes can be viewed as post-coordinated descriptors that provide, when combined, necessary and sufficient information to identify a particular (what could be called pre-coordinated) disease. For instance, the disease *Malignant neoplasm of lower-outer quadrant of breast* in ICD-10 corresponds to a malignant neoplasm arising from the lower outer quadrant of breast. More formally, this concept can be defined as a disease whose morphology is a malignant neoplasm and whose primary site is the lower outer quadrant of breast. In description logics, this can be stated as:

---

<sup>2</sup> <http://cbit.nci.nih.gov/evs-download>



*Malignant neoplasm of lower outer quadrant of breast*

$\equiv Disease \sqcap \exists has\_morphology.Malignant\ neoplasm$

$\sqcap \exists has\_primary\_site.Lower\ outer\ quadrant\ of\ breast$

In addition, because of its expressivity, ICD-O-3 provides finer-grained information about the morphology of disease than ICD-10 does. For instance, an adenocarcinoma arising from the lower-outer quadrant of breast is not available within ICD-10 because it is too specific. However, an adenocarcinoma being a type of malignant neoplasm, an adenocarcinoma arising from the lower-outer quadrant of breast is a type of malignant neoplasm arising from the lower outer quadrant of breast (which is a coarser grained concept that exists in ICD-10), formally equivalent to:

*Disease  $\sqcap \exists has\_morphology.Adenocarcinoma \sqcap \exists has\_primary\_site.Lower\ outer\ quadrant\ of\ breast$*

$\sqsubseteq Malignant\ neoplasm\ of\ lower\ outer\ quadrant\ of\ breast$

### **Adapting the defined pattern to the NCIt**

The NCIt provides both topographies, diseases and (even if not separately treated) morphologies. Fine-grained diseases, such as *Ductal breast carcinoma*, are also available. In order to address integration of diseases (ICD-10) with topographies and morphologies (ICD-O-3), the relationship *disease\_has\_primary\_anatomic\_site* is of particular interest. It is defined as “*A role used to relate a disease to the anatomical site where the originating pathological process is located. The domain and the range for this role are 'Disease, Disorder or Finding' and 'Anatomic Structure, System, or Substance'.*”. This relationship corresponds to the *has\_primary\_site* relationship defined above. In contrast, as no morphology axis is distinguished within the NCIt, it is not possible to find a relationship equivalent to the aforementioned *has\_morphology*. However, a disease is modeled as an anatomical specialization of a pre-coordinated morphology inside the NCIt.

Using the *disease\_has\_primary\_anatomic\_site* relationship and adapting the above description logics’ expression to the fact that no morphology axis is identified in the NCIt, we defined a pattern to formally describe relationships between diseases, morphologies and topographies within the NCIt. In description logics, this pattern is the following:

*Particular\_Disease  $\equiv$*

*Particular\_Morphology  $\sqcap \exists disease\_has\_primary\_anatomic\_site.Particular\_Topography$*

In the remainder of this section, we describe the methods implemented to use this pattern in order to systematically identify how pre-coordinated diseases can be mapped to combinations of a topography and a morphology within the NCIt. Our approach comprised two steps: the identification of morphologies and the search for mappings between diseases and topography-morphology combinations.

### Identifying morphologies within the NCIt

As stated before, the NCIt does not distinguish diseases from morphologies. To enable disease description through a topography-morphology combination, it was necessary to identify which NCIt concepts are morphologies and which are “pre-coordinated diseases”. Within ICD-O-3, morphologies and topographies are coded independently with no restriction. Morphology being a neoplastic characteristic it is not sufficient for describing a tumor (because a tumor should have at least one morphology and one primary anatomic site). Morphologies and diseases should be disjoint concepts since a disease cannot be a morphology and *vice versa*. Even if some morphologies are known to describe diseases that arise from a particular organ (e.g., hepatocellular carcinoma arising from the liver), morphologies cannot have an arising primary site stated; otherwise they have to be considered as diseases (because combining a morphology and a topography). In the NCIt, we thus restricted the domain of the *disease\_has\_primary\_anatomic\_site* relationship so that it cannot be a morphology. Formally, we wrote the following description logics’ expression:

$$\text{Morphology} \equiv \text{Neoplasm by morphology} \sqcap \forall \text{disease\_has\_primary\_anatomic\_site}.\perp$$

We executed the corresponding DL-Query and all extracted concepts were considered as morphologies. The latter were used to build topography-morphology combinations in the next step.

### Searching for topography-morphology combinations mapped to diseases within the NCIt

Based on the defined pattern, we implemented and executed the following algorithm:

- For each (*Morphology* identified  $\rightarrow$  [*Morphology*])
  - For each (*Anatomic Structure, System, or Substance* identified  $\rightarrow$  [*Topography*])
    - build [*expression*] of the form:  
[*Morphology*]  $\sqcap \exists \text{disease\_has\_primary\_anatomic\_site}.[*Topography*]$
    - recover equivalent classes of [*expression*]
    - recover direct subclasses of [*expression*]

The equivalent classes and direct subclasses retrieved were considered as diseases that could be mapped to the corresponding expressions (*i.e.*, topography-morphology combinations).

### Filtering the most relevant within resulting expressions

Searching for direct subclasses of the generated expressions is a mean to identify additional mappings to diseases. However, this approach may retrieve precise diseases mapped to expressions involving imprecise topographies or morphologies (e.g. *Adenocarcinoma of the gastroesophageal may be subclass of expressions involving both Adenocarcinoma and Carcinoma*). In order to address this issue, we filtered the obtained mappings so that only the most relevant expression mapped to a specific disease was kept.

*For each disease retrieved by more than one expression, we thus selected the expression involving the most precise morphology and the most precise topography as follows:*

- *If a disease is associated with a morphology and its superclasses then the morphology is selected as being the most precise and its superclasses are ignored.*
- *For topography selection:*
  - *If a disease is associated with a topography and its superclass's then the topography is selected as being the most precise and its superclasses are excluded.*
  - *If a disease is associated with a topography and classes on which it is located (according to the `anatomic_structure_has_location` relationship (e.g. `Duodenum` and `Epigastric Region`), then the topography is selected as being the most precise and classes on which it is located are excluded.*
  - *If a disease is associated with a topography and classes of which it is a part (according to the `anatomic_structure_is_physical_part_of` relationship (e.g. `Ascending Colon` and `Colon`), then the topography is selected as being the most precise and classes are excluded.*

## Evaluations

In order to assert that NCIt concepts mapped to ICD-O-3 morphologies have a formal description that respects the definition of ICD-O-3 morphologies, we evaluated how many NCIt concepts which are explicitly mapped to an ICD-O-3 morphology code within the NCIt were not retrieved as morphologies by our approach. We also evaluated how our method was able to map expressions (combining a topography and a morphology) to diseases. Toward this end, we compared the number of equivalent classes, direct subclasses and direct subclasses involving only the most relevant expression, which were obtained. As our aim was to find a unique expression as being equivalent to a unique disease, we reported the cardinalities of the expression-disease mappings that were retrieved.

## Results

### Morphologies identification and built expressions

The method proposed for identifying morphologies retrieved 632 distinct NCIt concepts. Among these concepts, 271 were mapped to at least one ICD-O-3 code within the NCIt. It is noteworthy that 941 (88%) of the NCIt concepts mapped to an ICD-O-3 code within the NCIt were not retrieved by our method because they all have a `disease_has_primary_anatomic_site` relationship associated with a topography (which is inappropriate for describing a morphology).

Combining the 632 identified morphologies with the 6,719 topographies available in the NCIt, 4,246,408 expressions were built. The corresponding DL-queries were executed in order to retrieve mappings between expressions and diseases within the NCIt.

### Mappings between expressions and diseases

Table 2 reports the number of concepts involved in mappings retrieved by each of the three approaches (equivalent classes of expressions, subclasses of expressions and subclasses of expressions involving only the most relevant

expression). Using equivalences, 12 pre-coordinated diseases were retrieved as equivalent to 12 expressions (Table 3). These statements involved 12 distinct topographies combined with 9 distinct morphologies. Each single pre-coordinated disease was equivalent to a single expression. Some of these equivalent statements looked unclassifiable (*i.e.*, *Parapharyngeal meningioma* arising from *head and neck*, which does not seem to be precise enough).

Table 2: Number of concepts involved in mappings retrieved according to the approach used.

|                           | Equivalences* | Subclasses** | Subclasses/Filtering*** |
|---------------------------|---------------|--------------|-------------------------|
| 7. Distinct diseases      | 8. 12         | 9. 2,231     | 10. 2,219               |
| 11. Distinct expressions  | 12. 12        | 13. 8,979    | 14. 2,228               |
| 15. Distinct topographies | 16. 12        | 17. 422      | 18. 327                 |
| 19. Distinct morphologies | 20. 9         | 21. 362      | 22. 323                 |

\*Equivalences correspond to diseases that are retrieved as being equivalent to expressions.

\*\*Subclasses correspond to diseases that are retrieved as being subclasses of expressions.

\*\*\*Subclasses/Filtering correspond to diseases that are retrieved as being subclasses of expressions involving only the most relevant topographies and morphologies.

When considering also subclass statements, 2,231 additional diseases were mapped to 8,979 expressions. Keeping only the most relevant expressions reduced significantly the number of distinct expressions mapped to diseases. The majority of the 95 topographies excluded by the filtering approach correspond to generic anatomic concepts (*Organ*, *Organ system*, *Other anatomic concept*), *body parts* (*Respiratory system part*, *Reproductive system part*) or imprecise concepts (*Cerebral white matter*, *connective and soft tissue*, *Ganglion*, *Large intestine*), which are not appropriate to describe precisely a given disease. The 39 morphologies excluded by the filtering approach were systemically related to diseases involving more precise morphologies. Compared with the subclass statements alone, the filtering approach excluded 12 diseases. The latter are the 12 diseases identified with the equivalence approach (because equivalent statements involved the most precise morphologies and topographies, thus confirming that our filtering approach was suitable).

Table 3: Equivalent statements identified for the 12 pre-coordinated diseases retrieved as equivalent to 12 expressions

| 23. Disease                      | 24. Expression                                                                                      |
|----------------------------------|-----------------------------------------------------------------------------------------------------|
| 25. Intravascular angioleiomyoma | 26. Angioleiomyoma $\sqcap$<br>$\exists$ disease_has_primary_anatomic_site.Vascular system          |
| 27. Abdominal fibromatosis       | 28. Deep fibromatosis/Desmoid tumor $\sqcap$<br>$\exists$ disease_has_primary_anatomic_site.Abdomen |
| 29. Lumbosacral lipoma           | 30. Deep lipoma $\sqcap$ $\exists$ has_primary_site.Lumbosacral region                              |
| 31. Tendon sheath lipoma         | 32. Deep lipoma $\sqcap$ $\exists$ disease_has_primary_anatomic_site.Tendon                         |
| 33. Parapharyngeal               | 34. Ectopic meningioma $\sqcap$                                                                     |

|                                       |                                                                                                                     |
|---------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| <i>meningioma</i>                     | $\exists$ <i>disease_has_primary_anatomic_site.Head and neck</i>                                                    |
| 35. <i>Fascial fibrosarcoma</i>       | <i>Fibrosarcoma</i> $\sqcap \exists$ <i>disease_has_primary_anatomic_site.Fascia</i>                                |
| 36. <i>Intravenous leiomyomatosis</i> | 37. <i>Leiomyomatosis</i> $\sqcap \exists$ <i>disease_has_primary_anatomic_site.Vein</i>                            |
| 38. <i>Neuroblastic tumor</i>         | 39. <i>Primitive neuroectodermal tumor</i> $\sqcap \exists$ <i>disease_has_primary_anatomic_site.Nervous system</i> |
| 40. <i>Plantar fibromatosis</i>       | 41. <i>Superficial fibromatosis</i> $\sqcap \exists$ <i>disease_has_primary_anatomic_site.Plantar region</i>        |
| 42. <i>Penile fibromatosis</i>        | 43. <i>Superficial fibromatosis</i> $\sqcap \exists$ <i>disease_has_primary_anatomic_site.Penis</i>                 |
| 44. <i>Sacrococcygeal teratoma</i>    | 45. <i>Teratoma</i> $\sqcap \exists$ <i>disease_has_primary_anatomic_site.Coccyx</i>                                |
| 46. <i>Gonadal teratoma</i>           | 47. <i>Teratoma</i> $\sqcap \exists$ <i>disease_has_primary_anatomic_site.Gonad</i>                                 |

#### Cardinalities of the retrieved expression-disease mappings

Table 4 reports the number of mappings obtained between expressions and diseases according to their cardinality and the approach used. Filtering relevant expressions reduced drastically N-aired cardinalities and retrieved a majority of 1-1 mappings between expressions and diseases. For instance, *Adenosquamous breast carcinoma* was initially mapped to two expressions (according to the “Subclasses” approach):

1. *Adenosquamous carcinoma*  $\sqcap \exists$  *disease\_has\_primary\_anatomic\_site.Breast*
2. *Adenosquamous carcinoma*  $\sqcap \exists$  *disease\_has\_primary\_anatomic\_site.Organ*

Because *Breast* is more precise than *Organ*, the filtering approach was able to select only the first expression.

Table 4: Number of mappings between expressions and diseases according to their cardinality and the approach used.

| 48.                                   | 49. Equivalences <sup>+</sup> | 50. Subclasses <sup>++</sup> | 51. Subclasses/Filtering <sup>+++</sup> |
|---------------------------------------|-------------------------------|------------------------------|-----------------------------------------|
| 52. 1 expression - 1 disease (1-1)*   | 53. 12                        | 54. 54                       | 55. 1,656                               |
| 56. 1 expression - N disease (1-N)    | 57. 0                         | 58. 6,028                    | 59. 398                                 |
| 60. N expressions - 1 disease (N-1)   | 61. 0                         | 62. 17                       | 63. 290                                 |
| 64. N expressions - N disease (N-N)** | 65. 0                         | 66. 14,034                   | 67. 218                                 |
| <b>Total</b>                          | <b>12</b>                     | <b>20,133</b>                | <b>2,562</b>                            |

\*1-1 means that for one expression, one disease was retrieved and this disease was not retrieved by another expression.

\*\*N-N means that for one expression, N diseases were retrieved and that at least one of these diseases was retrieved by another expression.

<sup>+</sup> Equivalences correspond to diseases that are retrieved as being equivalent to expressions.

<sup>++</sup> Subclasses correspond to diseases that are retrieved as being subclasses of expressions.

<sup>+++</sup> Subclasses/Filtering correspond to diseases that are retrieved as being subclasses of filtered expressions depending on the depth of the topographies and morphologies involved.

## Discussion

### Choice of the NCIt

In the biomedical field, other description logics-based terminologies exist. Specifically, SNOMED CT<sup>®</sup> provides not only topography, morphology and disease dimensions but also implements relationships between these concepts. However, the NCIt is specific to the oncology field and provides useful knowledge related to neoplasm diseases. In addition, it is freely and easily accessible. Finally, the NCIt “*Terms of use*” allows modifications and adaptations of its contents for specific purposes with no restriction. On the other hand, SNOMED CT has a much more restrictive affiliate license agreement and it is not easily accessible for countries which are not members of the IHTSDO<sup>3</sup>. In addition, it has been shown that SNOMED CT’s formal representation suffers from the same flaws [20,21] as the NCIt and has to be used cautiously while needing logic-based reasoning. Thus, a similar evaluation could be carried out over SNOMED CT in order to estimate if it could be useful for integrating disease terminologies in oncology.

### Limitations of the NCIt for integration purposes

The NCIt provides, on the one hand, pre-coordinated concepts describing cancer diseases and, on the other hand, simple concepts describing the tumor topography or the tumor morphology. It also provides relationships which are involved in topography-morphology combinations which are expected to be equivalences of pre-coordinated diseases. Its formal representation and the availability of an OWL version enable reasoning and the implementation of DL-queries. However, some intrinsic characteristics prevent its direct use for the integration of cancer disease terminologies: (i) the absence of distinction between morphologies and diseases and (ii) equivalences involving other axes than topography and morphology.

As the NCIt does not make any distinction between morphologies and diseases, we proposed a specific method for identifying the concepts that correspond to morphologies. However, the proposed method failed to recover 88% of the NCIt concepts which are explicitly mapped to ICD-O-3 morphology codes. In the NCIt, when the tumor morphology (mapped to an ICD-O-3 code) is known to have a specific primary site, it actually has a *disease\_has\_primary\_anatomic\_site* relationship with a topography. For example, the concept *Invasive ductal carcinoma, not otherwise specified*:

- has two corresponding ICD-O-3 morphology codes (8500/3 and 8521/3);
- is a subclass of the expression  $\exists$  *disease\_has\_primary\_anatomic\_site.Breast*.

These two statements are in contradiction because a tumor morphology does not entail a particular primary site from the ICD-O-3 point of view (this information is actually stated by the topographical axis of ICD-O-3). In [14],

---

<sup>3</sup> <http://www.ihtsdo.org/>

*infiltrating duct carcinoma* (8500/3) is used as an example: “[...] topography code C50.\_ (*Breast*) is added to the morphology term ‘infiltrating duct carcinoma’, [...]. However, if the term ‘infiltrating duct carcinoma’ is used for a primary carcinoma arising in the pancreas, coders should ignore the suggested breast topography code and assign the correct code, C25.9 (*pancreas, NOS*) instead”. Because its formal definition entails a primary site statement within the NCIt, our method did not retrieve (properly) *Invasive Ductal Carcinoma, Not Otherwise Specified* as being a tumor morphology. Thus, this inappropriate mapping stated in the NCIt may lead to semantic integration issues. Indeed, it is not possible to correctly represent an infiltrating duct carcinoma of pancreas coded in ICD-O-3 because it will automatically be reported as arising from the breast, which is not systematically true. Thus, the 941 NCIt concepts, which are mapped to ICD-O-3 codes although they have a *disease\_has\_primary\_anatomic\_site* with *topographies*, should be reviewed and corrected because their formal representation is not compliant with the definition of ICD-O-3 morphologies. This result emphasizes the need to distinguish morphologies and diseases within the NCIt which may help to correct the “*is\_a* overloading” problem as pointed in [17].

Only 12 topography-morphology combinations have been retrieved as equivalent to pre-coordinated diseases. Instead of defining simple equivalences based on topography and morphology, the NCIt often describes diseases according to other axes. For example, *Breast carcinoma* is defined as follows:

*Breast carcinoma*

≡ *Common carcinoma* ⊓ *Malignant breast neoplasm*

⊓ ∃ *disease\_has\_primary\_anatomic\_site.Breast*

⊓ ∃ *disease\_has\_normal\_tissue\_origin.Mammary epithelium*

As a result, a necessary and sufficient condition for a tumor to be a *Breast carcinoma* depends on information that is not available within ICD-O-3 axes (normal tissue origin is not available in ICD-O-3). Thus, for a large majority of diseases, it is not possible to identify an equivalent topography-morphology combination, which is another barrier for integration purposes.

### Implemented methods to address these issues

We have implemented a method for identifying morphologies within the disease axis of the NCIt. Our approach takes into account the ICD-O-3 definition of morphologies and uses the formal representation of the NCIt to identify concepts conforming to this definition. As stated aforementioned, our method could be used to identify probable incorrect mappings between ICD-O-3 morphologies and NCIt concepts.

In order to identify supplementary topography-morphology combinations related to pre-coordinated diseases, we extended our initial approach to recover subclasses of the corresponding expressions. We were able to filter topography-morphology combinations from non-relevant ones and to recover the corresponding pre-coordinated disease. This approach led to a large majority of non 1-1 relationships (Table 3). Indeed, many topography-morphology combinations were linked to multiple pre-coordinated diseases and *vice versa*. For example, *Adenocarcinoma of the gastroesophageal junction* was retrieved as a direct subclass of 16 expressions, including the following:

- *Adenocarcinoma*  $\sqcap \exists \text{disease\_has\_primary\_anatomic\_site. Gastroesophageal junction}$
- *Carcinoma*  $\sqcap \exists \text{disease\_has\_primary\_anatomic\_site. Gastroesophageal junction}$
- *Carcinoma*  $\sqcap \exists \text{disease\_has\_primary\_anatomic\_site. Gastrointestinal system part}$
- *Glandular cell neoplasm*  $\sqcap \exists \text{disease\_has\_primary\_anatomic\_site. Gastroesophageal junction}$
- *Glandular cell neoplasm*  $\sqcap \exists \text{disease\_has\_primary\_anatomic\_site. Gastrointestinal system part}$

The first expression involves the correct description of the targeted pre-coordinated disease while the other expressions provide insufficient information. However, all these expressions were retrieved as being a direct superclass of *Adenocarcinoma of the gastroesophageal junction*. In order to address this issue, we implemented a method to automatically select the most relevant combination. This approach used the NCIt graph to search for the most precise topographies and morphologies involved within multiple candidate expressions. Our method reduced significantly imprecise concepts without removing the identified disease, and produced successfully a large majority of 1-1 diseases-expressions mappings.

## Perspectives

This work confirms that logic-based reasoning over the NCIt in the context of a specific use case may lead to inconsistencies and that the OWL-DL version of the NCIt should be used cautiously. However, this resource may be helpful in order to build a formal model for integrating disease terminologies. Because our methods are fully automatic, there is a need for manual reviewing the identified disease-expression mappings. The evaluation process should consist in reviewing firstly equivalences, then filtered subclasses and, for the remaining diseases not mapped at the previous steps, the subclasses without filtering relevant combinations. Mappings must be assessed between ICD-O-3 and the NCIt and between ICD-10 and the NCIt. We are currently building a method based on the NCI Metathesaurus [22] to identify possible mappings that need to be manually reviewed.

## Conclusion

The NCIt is a rich knowledge resource. Its structure and the relationships available may help to bridge disease terminologies taking into account their structural and granularity heterogeneities. We have tried to extract concepts corresponding to topographies and morphologies involved in ICD-O-3 and concepts related to pre-coordinated diseases, which could be linked to ICD-10. These concepts can be retrieved and although morphology is not formally described, we have shown that it is possible to separate it from pre-coordinated diseases. In contrast, the formal description of relationships between these axes does not suit our specific usage. Indeed, the NCIt does not provide a morphological axis and does not describe diseases depending only on a topography-morphology combination. There is a need to build an ontology which integrates concepts manipulated by the disease terminologies in oncology and which formally describes their relationships. Although the OWL-DL version of the NCIt does not enable a straightforward solution, it may help to identify how involved concepts are related. We have developed an extraction method which can generate “candidate expressions”. However, these expressions have to be



reviewed in order to formally define links which actually exist between topography, morphology and disease in oncology. Moreover, while building a formal model for the integration of disease terminologies mapping identified concepts with the NCIt remains essential in enable the use of the rich knowledge related to diseases available within the NCIt.

## References

- [1] T. B. Murdoch et A. S. Detsky, « The inevitable application of big data to health care », *JAMA J. Am. Med. Assoc.*, vol. 309, n° 13, p. 1351-1352, 2013 apr.
- [2] H. U. Prokosch et T. Ganslandt, « Perspectives for medical informatics. Reusing the electronic medical record for clinical research », *Methods Inf. Med.*, vol. 48, n° 1, p. 38-44, 2009.
- [3] C. Safran, M. Bloomrosen, W. E. Hammond, S. Labkoff, S. Markel-Fox, P. C. Tang, D. E. Detmer, et Expert Panel, « Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper », *J. Am. Med. Inform. Assoc. JAMIA*, vol. 14, n° 1, p. 1-9, 2007 feb.
- [4] T. Botsis, G. Hartvigsen, F. Chen, et C. Weng, « Secondary Use of EHR: Data Quality Issues and Informatics Opportunities », *AMIA Summits Transl. Sci. Proc.*, vol. 2010, p. 1-5, 2010 mar.
- [5] C. Barton, C. Kallem, P. Van Dyke, D. Mon, et R. Richesson, « Demonstrating “Collect once, Use Many” – Assimilating Public Health Secondary Data Use Requirements into an Existing Domain Analysis Model », *AMIA. Annu. Symp. Proc.*, vol. 2011, p. 98-107, 2011.
- [6] A. El Fadly, B. Rance, N. Lucas, C. Mead, G. Chatellier, P.-Y. Lastic, M.-C. Jaulent, et C. Daniel, « Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform », *J. Biomed. Inform.*, vol. 44 Suppl 1, p. S94-102, 2011 dec
- [7] Black RJ, Simonato L, Storm HH, Démaret E, International Agency for Research on Cancer. Automated data collection in cancer registration. Lyon: IARC; 1998.
- [8] Jouhet V, Defossez G, Ingrand P. Automated Selection of Relevant Information for Notification of Incident Cancer Cases within a Multisource Cancer Registry. *Methods Inf Med* 2013;52. doi:10.3414/ME12-01-0101.
- [9] S. Tognazzo, B. Emanuela, F. A. Rita, G. Stefano, M. Daniele, S. C. Fiorella, et Z. Paola, « Probabilistic classifiers and automated cancer registration: An exploratory application », *J. Biomed. Inform.*, vol. 42, n° 1, p. 1-10, 2009 feb.
- [10] S. Tognazzo, A. Andolfo, E. Bovo, A. R. Fiore, A. Greco, S. Guzzinati, D. Monetti, C. F. Stocco, et P. Zambon, « Quality control of automatically defined cancer cases by the automated registration system of the Venetian Tumour Registry Quality control of cancer cases automatically registered », *Eur. J. Public Health*, vol. 15, n° 6, p. 657-664, 2005 jan.
- [11] P. Contiero, A. Tittarelli, A. Maghini, S. Fabiano, E. Frassoldi, E. Costa, D. Gada, T. Codazzi, P. Crosignani, R. Tessandori, et G. Tagliabue, « Comparison with manual registration reveals satisfactory completeness and efficiency of a computerized cancer registration system », *J. Biomed. Inform.*, vol. 41, n° 1, p. 24-32, 2008 feb.
- [12] Tagliabue G, Maghini A, Fabiano S, Tittarelli A, Frassoldi E, Costa E, et al. Consistency and accuracy of diagnostic cancer codes generated by automated registration: comparison with manual registration. *Popul Health Metr* 2006;4:10. doi:10.1186/1478-7954-4-10.
- [13] Organization WH. International statistical classification of diseases and related health problems. vol. 1. World Health Organization; 2004.
- [14] Fritz AG, World Health Organization. International classification of diseases for oncology: ICD-O. 2013.
- [15] Hartel F, Warzel DB, Covitz P. OWL/RDF/LSID Utilization in NCI Cancer Research Infrastructure. W3C Workshop Semantic Web Life Sci., 2004.
- [16] Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. *Methods Inf Med* 2005;44:498.
- [17] Kumar A, Smith B. Oncology ontology in the NCI thesaurus. *Artif. Intell. Med.*, Springer; 2005, p. 213–20.
- [18] S. Schulz, D. Schober, I. Tudose, et H. Stenzhorn, « The Pitfalls of Thesaurus Ontologization – the Case of the NCI Thesaurus », *AMIA. Annu. Symp. Proc.*, vol. 2010, p. 727-731, 2010.
- [19] NCI Thesaurus (NCIt) - EVS - National Cancer Institute - Confluence Wiki n.d. [https://wiki.nci.nih.gov/display/EVS/NCI+Thesaurus+\(NCIt\)](https://wiki.nci.nih.gov/display/EVS/NCI+Thesaurus+(NCIt)) (accessed March 3, 2015).

- [20] Héja G, Surján G, Varga P. Ontological analysis of SNOMED CT. BMC Med Inform Decis Mak 2008;8:S8. doi:10.1186/1472-6947-8-S1-S8.
- [21] Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies. Artif Intell Med 2007;39:183. doi:10.1016/j.artmed.2006.12.003.
- [22] NCI Metathesaurus n.d. <http://ncimeta.nci.nih.gov/ncimbrowser/> (accessed March 11, 2015)



## *Serment d'Hippocrate*

Au moment d'être admise à exercer la médecine, je promets et je jure d'être fidèle aux lois de l'honneur et de la probité.

Mon premier souci sera de rétablir, de préserver ou de promouvoir la santé dans tous ses éléments, physiques et mentaux, individuels et sociaux.

Je respecterai toutes les personnes, leur autonomie et leur volonté, sans aucune discrimination selon leur état ou leurs convictions. J'interviendrai pour les protéger si elles sont affaiblies, vulnérables ou menacées dans leur intégrité ou leur dignité. Même sous la contrainte, je ne ferai pas usage de mes connaissances contre les lois de l'humanité.

J'informerai les patients des décisions envisagées, de leurs raisons et de leurs conséquences.

Je ne tromperai jamais leur confiance et n'exploiterai pas le pouvoir hérité des circonstances pour forcer les consciences.

Je donnerai mes soins à l'indigent et à quiconque me les demandera. Je ne me laisserai pas influencer par la soif du gain ou la recherche de la gloire.

Admise dans l'intimité des personnes, je tairai les secrets qui me seront confiés. Reçue à l'intérieur des maisons, je respecterai les secrets des foyers et ma conduite ne servira pas à corrompre les mœurs.

Je ferai tout pour soulager les souffrances. Je ne prolongerai pas abusivement les agonies. Je ne provoquerai jamais la mort délibérément.

Je préserverai l'indépendance nécessaire à l'accomplissement de ma mission.

Je n'entreprendrai rien qui dépasse mes compétences. Je les entretiendrai et les perfectionnerai pour assurer au mieux les services qui me seront demandés.

J'apporterai mon aide à mes confrères ainsi qu'à leurs familles dans l'adversité.

Que les hommes et mes confrères m'accordent leur estime si je suis fidèle à mes promesses; que je sois déshonorée et méprisée si j'y manque.



## RESUME

---

**CONTEXTE :** Plusieurs terminologies sont utilisées en routine pour coder les diagnostics de cancer. Pour pouvoir utiliser conjointement des informations issues de différentes sources, il faut être en mesure de mettre en correspondance les diagnostics semblables entre eux, codées à l'aide de ces différentes terminologies, c'est à dire de rattacher sous un même concept des codes différents, et ce sans modification ou perte d'information. Le NCI thesaurus (NCIt) a semblé être une terminologie de choix pour l'élaboration d'une terminologie pivot afin de représenter de manière consensuelle et centralisée les concepts des différentes terminologies diagnostiques en cancérologie. La CIM-O3 a été choisie pour débiter le travail de mapping vers cette terminologie cible. **OBJECTIF :** L'objectif était d'évaluer la possibilité de représenter les données diagnostiques d'un registre général des cancers de l'adulte grâce à des concepts du NCIt, par l'intermédiaire d'un mapping validé établi préalablement entre la CIM-O3 et le NCIt et des relations existant entre les concepts du NCIt. **METHODE :** Une première phase a consisté en l'élaboration d'un mapping de la CIM-O3 vers le NCIt. La seconde phase s'est attachée à évaluer la possibilité de représenter les diagnostics réels du Registre Général des Cancers de la Gironde (RGCG), codés en CIM-O3, avec des concepts du NCIt. Pour chaque phase, une validation manuelle par une équipe de médecins a été réalisée. **RESULTATS :** Au 31 janvier 2015, la validation par l'équipe de médecins avait permis d'obtenir 605 mappings valides d'un code CIM-O3 vers un code NCIt (40,3 % des 1500 codes CIM-O3). Il existait 21 694 tumeurs (52,20% des 41 553 tumeurs enregistrées par le RGCG durant la période d'inclusion) ayant un mapping complet de leur diagnostic. Au total, 7822 tumeurs (18,8% des 41 553 tumeurs du RGCG) avaient un diagnostic qui pointait vers un diagnostic NCIt pré-coordonné valide.

**Mots clés :** CIM-O3, NCI Thesaurus, mapping de terminologies

## ABSTRACT

---

**BACKGROUND :** Several terminologies are routinely used to encode cancer diagnosis. To conjoint use information coming from several sources, it is necessary to be able to map similar diagnoses, encoded with those different terminologies, ie to attach under one concept different codes, without change or loss of information. The NCI thesaurus (NCIt) seemed to be a good foundation to develop a pivot terminology, to consensually represent in a centralized way the concepts from different diagnostic terminologies in oncology. The ICDO was chosen to begin the mapping work towards this target terminology. **OBJECTIVE :** The objective was to evaluate the possibility of representing the diagnostic data from a general cancers register of adult through the NCIt concepts, using a validated mapping previously established between ICDO and NCIt and relationships between NCIt concepts. **METHOD :** The first phase involved developing a mapping from ICDO to NCIt. The second phase focused on evaluating the possibility of representing the actual diagnoses of General Cancer Registry of the Gironde (RGCG), coded ICDO, with the nCit precoordinated concepts. For each phase, a manual validation by a team of physicians was performed. **RESULTS :** On January 31, 2015, the validation by the team of doctors had yielded 605 valid mappings of an ICDO code to a NCIt code (40,3 % of the 1500 ICDO codes). There were 21 694 tumors (52,20% of the 41,553 tumors registrated by the RGCG during the inclusion period) with a complete mapping of their diagnosis. A total of 7822 tumors (18.8% of the 41,553 tumors registrated by the RGCG) had a diagnosis that pointed to a valid NCIt pre-coordinated diagnosis.

**Key words :** CIM-O3, NCI Thesaurus, terminology mapping