Guide d’annotation des entités oncologiques selon la Classification Internationale des Maladies pour l’Oncologie, 3ᵉ édition (CIM-O-3 / ICD-O-3)

Septembre 2025

Version 1.2

Auteur·ice·s:

Johann Pignat

Milena Vucetic

Mina Bjelogrlic

# Introduction

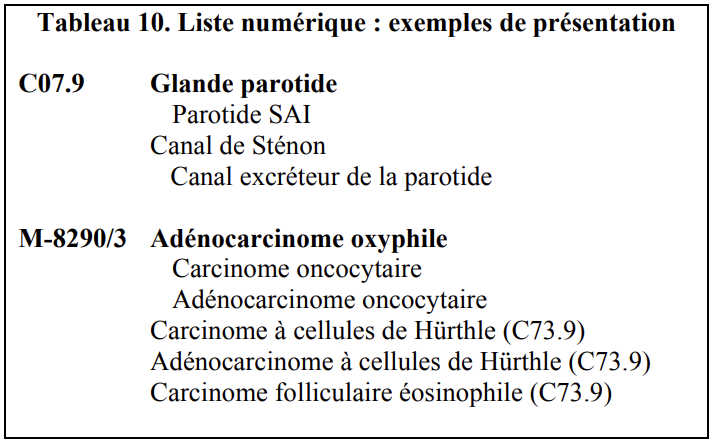
Ce document a pour objectif de fournir des règles claires et cohérentes pour l’annotation de narratifs cliniques (“texte libre”) avec la Classification Internationale des Maladies pour l’Oncologie, 3ᵉ édition (CIM-O-3 / ICD-O-3).

## “Entités oncologiques”

Nous nous intéressons ici à l’annotation “d’entités oncologiques”. Par “entité oncologique”, nous entendons tout fragment de texte qui décrit ou se rapporte à une tumeur, un cancer, une lésion néoplasique, qu’elle soit primaire ou secondaire (métastase), affirmée ou supposée. Ces entités sont décrites en pratique médicale par exemple dans les comptes rendus médicaux, les résultats d’examens, les lettres de consultation ou les comptes rendus opératoires. Ces entités incluent aussi bien la localisation anatomique de la tumeur (ex. : “sein gauche”), que la nature histologique (ex. : “carcinome canalaire infiltrant”), ou encore des précisions liées au grade ou à la différenciation cellulaire (ex. : “peu différencié”). L’objectif de l’annotation est d’identifier et de catégoriser systématiquement ces informations, afin de rendre les textes exploitables pour une indexation et une analyse automatisées.

## CIM-O-3/ICD-O-3

La CIM-O-3 est la référence internationale en matière de codage des cancers. Elle fournit un système standardisé composé de trois grands axes de description :

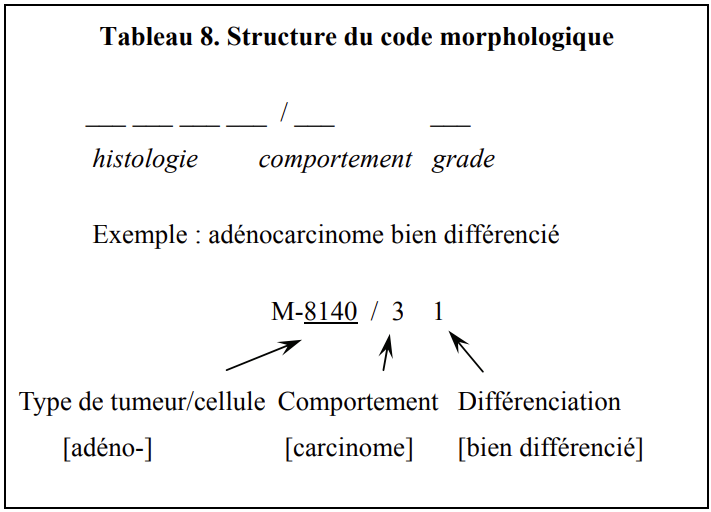
* La topographie : elle désigne le site anatomique où se situe la tumeur, en suivant la nomenclature de la CIM-10. Par exemple, le code C50 correspond au sein. Ce code « général » sera complété d’une précision de localisation, comme pour le code C50.1 qui représente par exemple la « région centrale du sein ».
* A white paper with black text

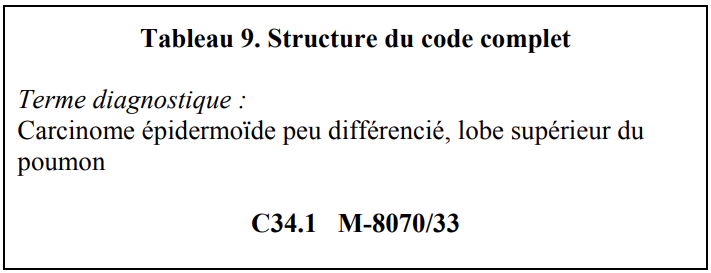
  AI-generated content may be incorrect.La morphologie : elle décrit la nature histologique de la tumeur (le type de cellules ou de tissus concernés) ainsi que son comportement biologique (bénin, malin, in situ, incertain). Les quatre premiers chiffres indiquent la description histologique de la tumeur, et le cinquième désigne le comportement de l’entité considérée (bénin, primaire, secondaire, incertain) Par exemple, le code 8500/3 correspond au “carcinome canalaire infiltrant, malin”.
* Le degré de différenciation (ou grade histologique) : il indique le niveau de ressemblance entre les cellules tumorales et les cellules normales, traduisant l’agressivité et le pronostic du cancer. Les valeurs typiques vont de “bien différencié” à “indifférencié”.

A white paper with black text

AI-generated content may be incorrect.Ce guide présente également une annotation composite qui englobe les différents composants d’une même expression oncologique pour correspondre au « code complet » décrit dans la CIM-O.

L’ensemble de ces axes permet de décrire de façon précise et normalisée tout cas de cancer. Dans le cadre de ce guide, ils serviront de référence pour l’annotation des entités oncologiques, afin que chaque mention dans un texte libre puisse être reliée à une catégorie CIM-O-3 pertinente.





# Entités à annoter

Ce guide d’annotation vise à l’annotation de 4 types d’expressions qui sont détaillées dans la suite de ce guide :

* Expressions « éléments » :
  + Expressions de morphologie tumorale
  + Expressions topographiques liées à une morphologie tumorale
  + Expressions de différenciation tumorale
* Expressions « complètes » réunissant une morphologie avec ses précisions de topographie et de différenciation

# Considérations pratiques

Pour la création des annotations, nous utilisons l’outil BRAT (Brat Rapid Annotation Tool). L’interface est simple et se manipule directement avec la souris. La marche à suivre est la suivante :

1. Sélectionner dans le texte l’expression à annoter à l’aide de la souris.
2. Choisir dans la fenêtre contextuelle qui s’affiche le label approprié parmi : morphologie, topographie, différenciation ou expression complète.
3. Renseigner dans le champ Notes le ou les codes ICD-O-3 correspondants. En cas de multiplicité, les codes doivent être séparés par un point-virgule “;”.

Cette procédure garantit que chaque entité est à la fois identifiée, catégorisée et liée à la nomenclature ICD-O-3.

A screenshot of a computer

AI-generated content may be incorrect.

# Règles d’annotations

Les règles d'annotation sont détaillées dans cette section, avec des exemples destinés à faciliter la compréhension des directives. Les règles d'annotation reprennent les bases établies pour l’annotation du corpus CANTEMIST et sont structurées selon les classes suivantes :

* Règles générales
* Règles positives
* Règles négatives
* Règles orthographiques
* Règles de conventions et de codage
* Exceptions

## Règles générales

### G1. Utilisation de ressources de connaissances externes

Si l'annotateur·ice a des doutes quant aux expression à annoter ou à leur codification en CIM-O malgré consultation de la documentation officielle, il est conseillé de consulter des ressources externes en français (par exemple, Wikipédia ou des sites web spécialisés. D'autres ressources en anglais facilitant la compréhension du terme peuvent également être consultées : Wikipédia, PubMED, d'autres terminologies médicales : SNOMED CT, ICD-10, ou d'autres bases de données pour dissiper tout doute. Les principales ressources de référence sont :

* Classification Internationale des Maladies pour l’Oncologie. Troisième Édition : <https://iris.who.int/server/api/core/bitstreams/1a774b8b-f31d-42ea-9ac3-8b7d9e6140a7/content>
* Manuel de la CIM-10, Version:2008 ou ultérieure : <https://icd.who.int/browse10/2008/fr>
* Abréviations médicales suisses : <https://abreviationsmedicales.ch/>
* SNOMED CT browser : <https://browser.ihtsdotools.org/?perspective=full&conceptId1=237617006>

### G2. Portée spécifique de l'annotation

Le processus d'annotation ne vise pas à annoter tous les mots inclus dans le cas clinique.

L'objectif est de trouver les entités nommées listées dans la section 2 - Entités à annoter, dont les détails sont disponibles plus bas.

### G3. Isolation de l’annotateur·ice

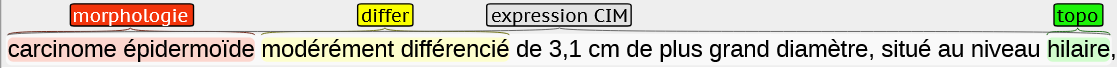
Un·e annotateur·ice ne doit pas demander d’avis à un autre durant le processus d’annotation afin de garantir une indépendance dans l’annotation et de permettre une étude informative de l’accord inter-annotateurs.

### G4. Révision des règles d’annotation

Si des cas particuliers d’expressions susceptibles d'être intéressants sont détectés et que les directives ne précisent pas leur annotation, ces cas doivent être signalés avec des exemples et une brève explication pour permettre d’affiner les règles d'annotation par la suite.

### G5. Expression complète

L’expression complète regroupe les éléments de morphologie, de topographie et de différenciation d’une même entité oncologique. L’annotation doit commencer au début du premier élément annoté et terminer à la fin du dernier élément annoté. Une expression complète correspond à exactement un élément morphologique et inversement.



### G6. Expressions multiples

Les expressions morphologiques et expressions complètes doivent être annotées à chaque apparition, même si le texte est exactement le même ; autrement dit, une même entité oncologique mentionnée plusieurs fois doit être annotée autant de fois qu’elle apparaît dans le texte. Une expression dont l’abréviation est présentée directement après (p.ex. « carcinome canalaire infiltrant (CCI) ») doit être annotée d’un seul bloc.

### G7. Expressions conditionnelles

Les expressions de topographie et de différenciation ne doivent être annotées que si elles peuvent être liées à une mention explicite d’une expression morphologique dans le texte. Les expressions topographiques non liées à une pathologie tumorale ne doivent donc pas être annotées.

### G8. Annotation minimale

Toute annotation complète doit couvrir au minimum un mot entier et être composée au minimum d’une expression morphologique et d’une expression complète. L’expression complète est superposée à toutes les expressions éléments d’une même expression. Les expressions de topographie et de différenciation ne sont pas obligatoires, mais doivent être annotées si elles sont présentes et reliées à une expression morphologique explicite.

### G9. Annotation fragmentée

Pour les annotations non complètes (morphologie, topographie, différenciation) il est possible de « fragmenter » afin d’annoter une expression séparée par une autre.

A close-up of a logo

AI-generated content may be incorrect.

### G9. Superposition des expressions

Il est interdit de superposer des annotations d’expressions non complètes. Il peut arriver cependant que deux expressions complètes se chevauchent. Les règles précédentes s’appliquent normalement.

A close-up of a logo

AI-generated content may be incorrect.

### G10. Limite des expressions

Il est possible que certains éléments d’une expression complète apparaissent dans les phrases suivant l’expression morphologique. Dans ce cas il est permis d’annoter des expressions à cheval sur plusieurs phrases, mais il convient que ce chevauchement ne pas dépasse pas deux phrases dans la mesure du possible, et ce pour des raisons de clarté des annotations.

### G11. Expression minimale

Une entité doit être reconnue par son expression la plus complète. Même si la CIM-O ne permet une granularité très importante, notamment au niveau topographique, il convient d’annoter l’expression la plus complète qui se réfère même si le code CIM-O pourrait être attribuée à une partie plus courte, à l’exception faite des précisions de latéralité (droite, gauche).

A close-up of a computer screen

AI-generated content may be incorrect.

### G12. Exclusion de la latéralité

En exception à la règle G11, la latéralité des entités topographiques est à exclure des annotations, étant donné que ces précisions ne sont pas couvertes par la CIM-O. La seule exception à cette règle est lorsque la latéralité fait partie d’une abréviation (p.ex. LID, lobe inférieur droit), conformément à la règle G8.

## Règles positives

### P1. Pathologie oncologique

Toute expression faisant référence à une potentielle pathologie oncologique doit être annotée, y compris les termes peu spécifiques, qui dans le contexte font référence à une (potentielle) pathologie oncologique (p.ex. « lésion », « masse », « nodule », « tumeur »).

### P2. Abréviations

Les abréviations peuvent être considérées comme des expressions si tant est que l’abréviation est reconnaissable ou que le contexte en permet une identification simple et sans ambiguïté (p.ex. LH, annoté uniquement dans le contexte du lymphome de Hodgkin et pas de *luteinizing hormone*).

### P3. Antécédents et anamnèse familiale

Toute mention d’une pathologie oncologique potentielle doit être annotée, même s’il s’agit d’une description d’antécédents ou d’anamnèse familiale. Tant que l’expression désigne une pathologie oncologique « tangible » (ou son absence, cf. P4), elle doit être annotée. Ceci exclut donc les expressions telles que « comité du cancer du sein », ou « recommandations dans les cas de cancer du côlon », qui ne doivent donc pas être annotées.

### P4. Expressions « négatives »

Les expressions complètes ou morphologique faisant parties d’une négation ou d’une absence « négatives » (p.ex. « absence de cancer » ou « pas de lésion suspectes ») doivent être annotées si elles répondent aux autres règles, à l’exception de l’expression « M0 ». Il en va de même pour les expressions indéterminées (p.ex. « suspicion de métastase », « compatible avec un carcinome »). Cette règle s’applique à l’expression morphologique de l’expression complète, à l’inverse, les expressions de topographie et de différenciation niées séparément ne doivent pas être annotées.

### P5. Progression, évolution de la maladie

Les expressions désignant une évolution de la maladie (p.ex. « progression de la maladie », « récidive tumorale ») doivent être annotées.

### P6. Métastases, M1, « stade IV »

Toutes les expressions désignant des métastases, et par extension l’expression « stade IV » (qui sous-entend la présence de métastases) doivent être annotées comme métastases. L’expression « M1 » peut faire dérogation à la règle G8 si elle fait partie d’une expression TNM plus complète. Il a été décidé que la métastase devait être une annotation séparée de la pathologie principale même si ces deux sont présentes dans la même expression.

A close-up of a word

AI-generated content may be incorrect.

### P7. Préfixes

Conformément à la règle G8, un mot composé d’une expression d’intérêt et d’un suffixe (p.ex. « micrométastatique », « périhilaire ») doit être annoté dans son entièreté. Cette règle ne s’applique que si le terme conserve son intérêt malgré le suffixe (p.ex. « paratumoral » ne sera pas annoté sans contexte, car il ne désigne pas directement l’entité oncologique).

## Règles négatives

### N1. Diagnostics différentiels

En complément à la règle P3, il ne faut pas annoter les énumérations de diagnostics différentiels avec les différentes possibilités diagnostiques (p.ex. « autres étiologies possibles : lymphome non-hodgkinien, tuberculose, nodules bénins »).

### N2. Expressions adjectivales

En général, il convient de ne pas annoter les expressions morphologiques si elles sont adjectives à un terme non lié à la morphologie (p.ex. « traitement tumoral »).

### N3. Expressions implicites

Les expressions qui représentent implicitement (ou par leur contexte uniquement) une pathologie tumorale ne doivent pas être annotées (p.ex. « le second est au niveau pulmonaire »).

### N4. Expressions de stade

À l’exception du « stade IV » et de « M1 » dans la classification TNM, les stades tumoraux ne doivent pas être annotés (p.ex. « stade IIa », « T4N0 »).

## Règles orthographiques

### O1. Autres langues

Les expressions pertinentes dans d'autres langues que le français doivent être annotés, quelle que soit la langue, et même s’il s’agit d’abréviations.

### O2. Erreurs typographiques et de conversion

Les mentions de morphologie néoplasique (qui respectent les règles générales d'annotation) et qui comportent des erreurs typographiques doivent être annotées. Cela inclut les mentions présentant des erreurs de conversion automatique, générées par des programmes de conversion de texte.

### O3. Espaces incorrects

Les mots séparés en plusieurs unités de texte par des espaces incorrectement placés doivent être annotés comme une seule entité.

### O4. Déterminants

Il ne faut pas annoter les déterminants qui introduisent une expression à annoter (p.ex. « de grade II », seul « grade II » sera annoté).

### O5. Ponctuation, espaces, parenthèses, apostrophes et apparentés

Il ne faut pas pousser l’annotation au-delà du dernier mot de l’expression pour capturer un point, une virgule, ou la fin d’un parenthèse qui aurait commencé dans l’expression (p.ex. « carcinome canalaire infiltrant (CCI) », seul sera annoté « carcinome canalaire infiltrant (CCI »). Il en va de même pour une apostrophe qui précède l’expression (p.ex. « l’atteinte tumorale », seule sera annotée « atteinte tumorale »).

## Règles de conventions et de codage

### C1. Codage des expressions-élément

Les expressions-élément (morphologie, topographie, différenciation) doivent être codée en assignant au mieux les codes définis par la CIM-O-3 pour les catégories respectives.

### C2. Codage des expressions complètes

Les expressions complètes doivent être codées en rassemblant les différents éléments de l’expression et se présenter au format « topographie morphologie+différenciation » (p.ex. « C50.9 8000/31 »). Si moins d’éléments sont disponibles le code d’une expression complète peut être composée au minimum d’un code morphologique (p.ex. « 8500/3 »).

### C3. Topographies usuelles

Si une topographie usuelle est présentée dans la CIM-O à côté d’une expression morphologique, elle peut être ajoutée au code de l’expression complète si elle est compatible avec l’expression même si elle n’apparaît pas textuellement. La précision du code (3ème chiffre) est alors « 9 » pour représenter que le site exact n’est pas défini (p.ex. « carcinome canalaire infiltrant » [C50.9 8500/3]).

### C4. Utilisation du 5ème chiffre de la morphologie (comportement)

Le 5ème chiffre du code morphologique doit être utilisé, même si le terme spécifique n'est pas répertorié dans la CIM-O. Si l'annotateur·ice détermine que le comportement diffère du comportement habituel indiqué dans la CIM-O, il peut le modifier selon les règles de comportement de la CIM-O. Le système de la CIM-O-3 permet en effet de modifier un code donné suivant si la pathologie oncologique concernée est considérée comme bénigne (0), maligne (3), in situ (2), de malignité ou de bénignité incertaine (1), ou « maligne, incertain si primitif ou métastatique » (9). Se référer à au tableau de l’introduction du présent document.

### C5. Codage des tumeurs mixtes

Conformément aux indication de la CIM-O, l’expression d'une tumeur mixte composée de plusieurs morphologies représentées dans la CIM-O-3 comme concepts distincts. En mode combiné, la déclaration sera notée comme une déclaration continue, et sera codée comme le type principal, ou le premier type mentionné ou encore le type général si le type principal ne peut pas être défini. (p.ex. « tumeur germinale mixte maligne (choriocarcinome 60 % + tératome mature » [9101/3]).

### C6. Utilisation du 6ème chiffre pour la différenciation cellulaire

Lorsque disponibles, les codes pré-coordonnés de la CIM-O-3 spécifiant le degré de différenciation cellulaire doivent être utilisés et ajoutés au

## Cas spéciaux

### S1. Adénopathies

Dans le cas spécifique des « adénopathies » et d’autres expressions dérivées (p.ex. « conglomérat adénopathique »), qui représentent à la fois une considération morphologique de pathologie potentiellement tumorale et une topographie (ganglions lymphatiques), il convient d’annoter « adénopathie » ou son terme dérivé comme une expression morphologique. Les précisions sur la localisation de l’adénopathie seront ensuite considérées comme expressions topographiques. De plus la codification de l’expression complète devra prendre en compte l’expression entière, et non les sous-éléments pour l’attribution du code (p.ex. « adénopathie [8000/1] thoracique [C76.1] » donnera le code complet [C77.1 8000/1].)

### S2. Topographie contenue dans la morphologie

Dans certains cas spécifiques de la CIM-O-3, une expression topographique est incluse à l’expression morphologique. Lorsque c’est le cas dans la CIM-O-3, il convient d’annoter le tout comme expression morphologique et d’y attribuer le code précisé par la CIM-O (p.ex. « tumeur bénigne du stroma gastrointestinal [8936/0]»).

# Références

Outil d’annotation et de visualisation brat : <https://brat.nlplab.org/>

CIM-O-3 : <https://iris.who.int/server/api/core/bitstreams/1a774b8b-f31d-42ea-9ac3-8b7d9e6140a7/content>

Guidelines CANTEMIST : <https://temu.bsc.es/cantemist/?p=4362>