# Interprétation de bonnes pratiques de codification médicale par du raisonnement à partir de cas - Application à la saisie de données pour les registres du cancer

Michael Schnell, Sophie Couffignal, Jean Lieber, Stéphanie Saleh, Nicolas Jay

▶ **To cite this version:**

Michael Schnell, Sophie Couffignal, Jean Lieber, Stéphanie Saleh, Nicolas Jay. Interprétation de bonnes pratiques de codification médicale par du raisonnement à partir de cas - Application à la saisie de données pour les registres du cancer. Journée I.A. et Santé, Jul 2018, Nancy, France. hal-01907088

**HAL Id: hal-01907088**
**https://hal.science/hal-01907088v1**

Submitted on 28 Oct 2018

**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.

# Interprétation de bonnes pratiques de codification médicale par du raisonnement à partir de cas — Application à la saisie de données pour les registres du cancer

Michael Schnell[1,2], Sophie Couffignal[1], Jean Lieber[2],
Stéphanie Saleh[1], Nicolas Jay[2,3]

[1] Department of Population Health, Luxembourg Institute of Health,
1A-B, rue Thomas Edison, L-1445 Strassen, Luxembourg firstname.lastname@lih.lu

[2] Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy
firstname.lastname@loria.fr

[3] Service d'évaluation et d'information médicales, Centre Hospitalier
Régional Universitaire de Nancy, Nancy, France n.jay@chru-nancy.fr

**Résumé** : Les registres du cancer jouent un rôle important dans la lutte contre le cancer. Afin d'obtenir des données de qualité et comparables, des standards de codification internationaux et maintes bonnes pratiques de codification médicale doivent être respectés. Cet article propose une approche d'aide aux encodeurs dans le cadre des registres du cancer et décrit l'application de raisonnement à partir de cas. Cet article détaille également une implémentation prototype de l'approche décrite et des résultats préliminaires.

**Mots-clés** : interprétation bonnes pratiques, raisonnement à partir de cas interprétatif, standards de codification, registres du cancer, aide à la saisie, aide à la décision.

## 1 Introduction

Il existe de nombreux registres du cancer dans le monde. Ces registres collectent des données sur les cas de cancer diagnostiqués et/ou traités dans une région donnée. Ces données servent au suivi du cancer (taux d'incidence, taux de survie, etc.) et à l'évaluation de la prise en charge des patients (diagnostic, traitement, etc.). Afin de fournir des données comparables, il faut respecter les définitions communes (par exemple des terminologies comme la classification internationale des maladies (CIM)) et les standards de codification (European Network of Cancer Registries and Tyczynski, Jerzy E and Démaret, D and Parkin, D Maxwell, 2003). Cependant, l'étendue et la complexité des standards compliquent la tâche des encodeurs.

L'objectif de ce projet de recherche est d'atténuer cette difficulté, en aidant les encodeurs et les experts de codification dans l'interprétation et l'application des bonnes pratiques de codification.

L'exemple suivant sert à illustrer une situation dans laquelle un encodeur peut avoir recours à un expert de codification pour l'encodage d'un cas de cancer. Le cas concerne une femme pour laquelle en 2016 une imagerie met en évidence une image en lâcher de ballons [1] dans le poumon droit. Un CT scan ne met en évidence aucune adénopathie médiastinale [2]. Une analyse histologique permet d'identifier la morphologie [3] de la tumeur comme étant un adénocarcinome, positif au test du marqueur TTF1. D'autres examens ont mis en évidence

1. Une image en lâcher de ballons denote la présence de nombreuses images nodulaires arrondies.

2. Une adénopathie est affection des nœuds (ganglions) lymphatiques superficiels et profonds se traduisant par une augmentation anormale de leur volume (cf. dictionnaire médical de l'académie de médecine).

3. La morphologie d'une tumeur décrit le type et le comportement des cellules tumorales.

un deuxième site tumoral au niveau des ovaires. Dans cette situation, un encodeur peut s'interroger sur la topographie de la tumeur primitive, c'est-à-dire la localisation initiale de la tumeur (poumon, ovaires ou autre ?). Dans le cadre du Registre National du Cancer du Luxembourg (RNC), un encodeur peut faire appel aux experts de codification du RNC pour répondre à sa question, en utilisant le système en ligne de tickets du RNC. Les experts de codification utilisent la description textuelle mise à disposition par les encodeurs pour fournir une solution, c'est-à-dire une réponse accompagnée d'un argumentaire.

La section 2 décrit une approche d'aide à la codification pour les registres du cancer et l'intégration du raisonnement à partir de cas (RAPC (Aamodt & Plaza, 1994)) dans cette approche. La section 3 présente un prototype et quelques résultats préliminaires. La section 4 conclut cet article et présente de futures pistes de recherche.

## 2 Raisonnement à partir de cas argumentatif

Cet article résume les travaux réalisés dans (Schnell *et al.*, 2017) et ajoute une description du prototype réalisé ainsi que quelques résultats préliminaires.

### 2.1 Préliminaires

Un cas $(\texttt{srce}, \texttt{sol}(\texttt{srce}))$ est composé de deux éléments : 1) une description du dossier du patient et une question, et 2) une solution.

Le dossier du patient contient toutes les informations extraites du dossier hospitalier du patient pertinentes pour répondre à la question. Cette liste d'informations varie en fonction du sujet et est définie par les experts de codification. Le dossier du patient est représenté par un graphe RDFS (Brickley & Guha, 2014). Les parties du corps et les morphologies utilisent des classes de l'ontologie SNOMED Clinical Terms [4]. Les figures 1 et 3 montrent un extrait d'un graphe RDFS pour le dossier patient de l'exemple fourni dans l'introduction.

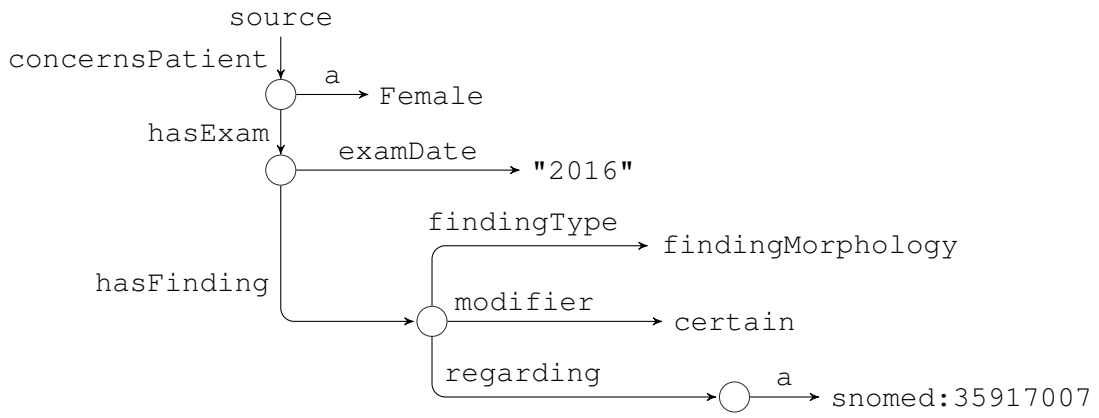


FIGURE *1 – Graphe RDFS partiel représentant une patiente et l'examen qui a permis d'identifier la morphologie de la tumeur. Les cercles représentent des nœuds anonymes. La ressource snomed:35917007 correspond à la classe SNOMED des adénocarcinomes.*

La question décrit le sujet (date d'incidence [5], topographie, nature de la tumeur, etc.). Dans l'exemple précédant, la question concerne la topographie de la tumeur. La solution contient la réponse à la question et les arguments les plus importants en faveur (**pros**) et en défaveur (**cons**) de la réponse.

4. https://bioportal.bioontology.org/ontologies/SNOMEDCT

5. La date d'incidence est la date du premier événement survenu qui a permis de poser le diagnostic de tumeur primitive.

Les arguments ont deux finalités. 1) Les arguments servent à expliquer la réponse aux encodeurs et servent de rappel pour les experts de codification. 2) Ils sont également utilisés dans l'étape de recherche de l'approche proposée.

Trois types d'arguments sont considérés : fort en faveur (strong pro), faible en faveur (weak pro) et faible en défaveur (weak con). La distinction entre un argument fort et faible se fait par rapport à son degré de certitude pour la réponse soutenue. Un argument fort est une justification suffisante pour conclure la réponse soutenue par l'argument. Contrairement, un argument faible n'est qu'une indication. A noter qu'il n'existe pas d'argument fort en défaveur. En effet, un tel argument serait une justification indiscutable pour exclure la réponse retenue pour le cas concerné. Formellement, un argument est une fonction `a` qui à un cas associe un Booléen et est enregistrée sous forme d'une requête SPARQL ASK. La figure 2 illustre un argument et sa formalisation en SPARQL.

`a(case) =`

```
ASK {
  case concernsPatient ?patient .

  ?patient hasExam ?exam_morpho .
  ?exam_morpho hasFinding ?finding .
  ?finding findingType findingTypeFindMorphology .
  ?finding modifier certain .
  ?finding regarding [ a snomed:35917007 ] .

  ?patient hasExam ?exam_ttf .
  ?exam_ttf hasFinding ?finding .
  ?finding findingType findingTypeFindTTF1Marker .
  ?finding modifier certain .
  ?finding present yes .
}
```

FIGURE *2 – Formalisation d'un argument* `a` *: un adénocarcinome TTF1 positif est en faveur d'une origine primitive pulmonaire. L'argument vérifie que la morphologie de la tumeur est de type adénocarcinome TTF1 positif. Cet argument s'applique à l'exemple de cas* `exmpl` *fourni dans l'introduction, autrement dit* `a(exmpl) = TRUE`*.*

### 2.2 Architecture globale

L'approche proposée utilise un cycle 4-R (*retrieve, reuse, revise, retrain*) adapté de (Aamodt & Plaza, 1994) et quatre conteneurs de connaissances (Richter & Weber, 2013) (*case base, domain knowledge, retrieval knowledge, adaptation knowledge*).

### 2.3 Recherche (retrieve)

L'approche proposée repose sur des arguments pour identifier des cas similaires. En effet, des réponses similaires devraient avoir des raisonnements similaires et donc les mêmes arguments devraient s'appliquer. Notre méthode vérifie l'applicabilité des arguments des cas sources pour le problème cible `tgt` et en déduit quel est le cas source le plus approprié pour résoudre `tgt`. La comparaison entre deux cas sources `i` et `j` repose sur trois critères, concernant 1) les arguments forts $\Delta^{\mathtt{s}}_{\mathtt{i,j}}$, 2) les arguments faibles $\Delta^{\mathtt{w}}_{\mathtt{i,j}}$ et 3) la similarité des dossiers patients $\Delta^{\mathtt{dist}}_{\mathtt{i,j}}$. Pour les arguments forts, le cas source avec le plus grand nombre d'arguments forts s'appliquant à `tgt` est le plus approprié. Pour les arguments faibles, une combinaison des arguments en faveur et en défaveur est considérée. Le cas source avec le plus d'arguments en faveur et le moins d'arguments en défaveur s'appliquant est le plus

approprié. Pour la similarité des dossiers patients, une distance d'édition (Bunke & Messmer, 1993) entre les graphes RDFS des dossiers est considéré. Le cas source avec le dossier patient le plus proche (c'est-à-dire la distance la plus petite) est le plus approprié. Ces trois critères sont évalués par ordre lexicographique, d'abord $\Delta^{\mathtt{s}}_{\mathtt{i,j}}$, puis $\Delta^{\mathtt{w}}_{\mathtt{i,j}}$ et finalement $\Delta^{\mathtt{dist}}_{\mathtt{i,j}}$.

### 2.4 Réutilisation (reuse)

Lorsque le cas source le plus approprié a été identifié, la solution associée est copiée pour résoudre $\mathtt{tgt} : \mathtt{sol(tgt)} := \mathtt{sol(srce)}$. Les arguments du cas source qui ne s'appliquent pas sont supprimés de la solution copiée.

### 2.5 Révision et apprentissage (revise and retain)

Le nouveau cas $(\mathtt{tgt}, \mathtt{sol(tgt)})$ peut être révisé par un expert de codification, afin de modifier la réponse, les arguments et/ou le dossier patient. Par exemple, un expert peut supprimer des informations inutiles du dossier patient par rapport à la question posée, afin de généraliser le cas et de mettre en valeurs les informations pertinentes pour l'explication de la réponse. Dans ce scénario, $(\mathtt{tgt}, \mathtt{sol(tgt)})$ est remplacé par $(\mathtt{tgt}', \mathtt{sol(tgt}'))$, tel que $\mathtt{tgt}'$ est plus général que $\mathtt{tgt}$. $(\mathtt{tgt}', \mathtt{sol(tgt}'))$ est un cas généralisé avec un périmètre plus large que $(\mathtt{tgt}, \mathtt{sol(tgt)})$ (Maximini *et al.*, 2003).

## 3 Prototype et résultats préliminaires

Le prototype réalisé dans le cadre du RNC sert de système en ligne de tickets, où les encodeurs peuvent poser des questions et les experts de codification peuvent répondre à des questions. Il guide les encodeurs dans la structuration de leurs questions, facilitant ainsi par la suite la recherche de questions similaires pour le RNC et les experts de codification. Pour les questions concernant la topographie, le prototype propose une solution provisoire, déduite à l'aide de la méthode présentée dans cet article. Toutes les réponses sont révisées par les experts de codification.

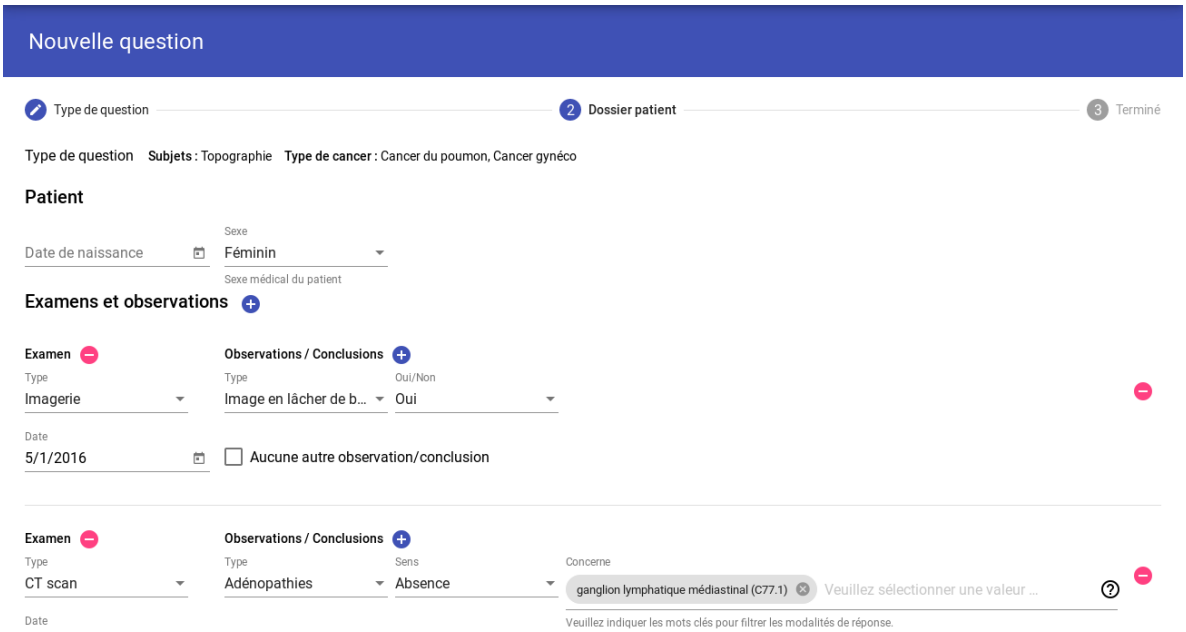


FIGURE 3 – *Extrait de l'interface de saisie d'un cas. En fonction des sujets choisis, l'encodeur dispose de différents items à remplir pour décrire le dossier du patient.*

Le prototype se présente sous la forme d'une application web, avec un client lourd de type Single Page Application construit à l'aide d'Angular [6] et une API de type REST écrite en Go [6] avec la libraire Gin [6]. Les données sont stockées dans un triple store Apache Jena [7] et sont exposés par un serveur SPARQL Apache Fuseki [7]. Les figures 3 et 4 montrent des captures d'écrans du prototype réalisé pour l'exemple décrit dans l'introduction.

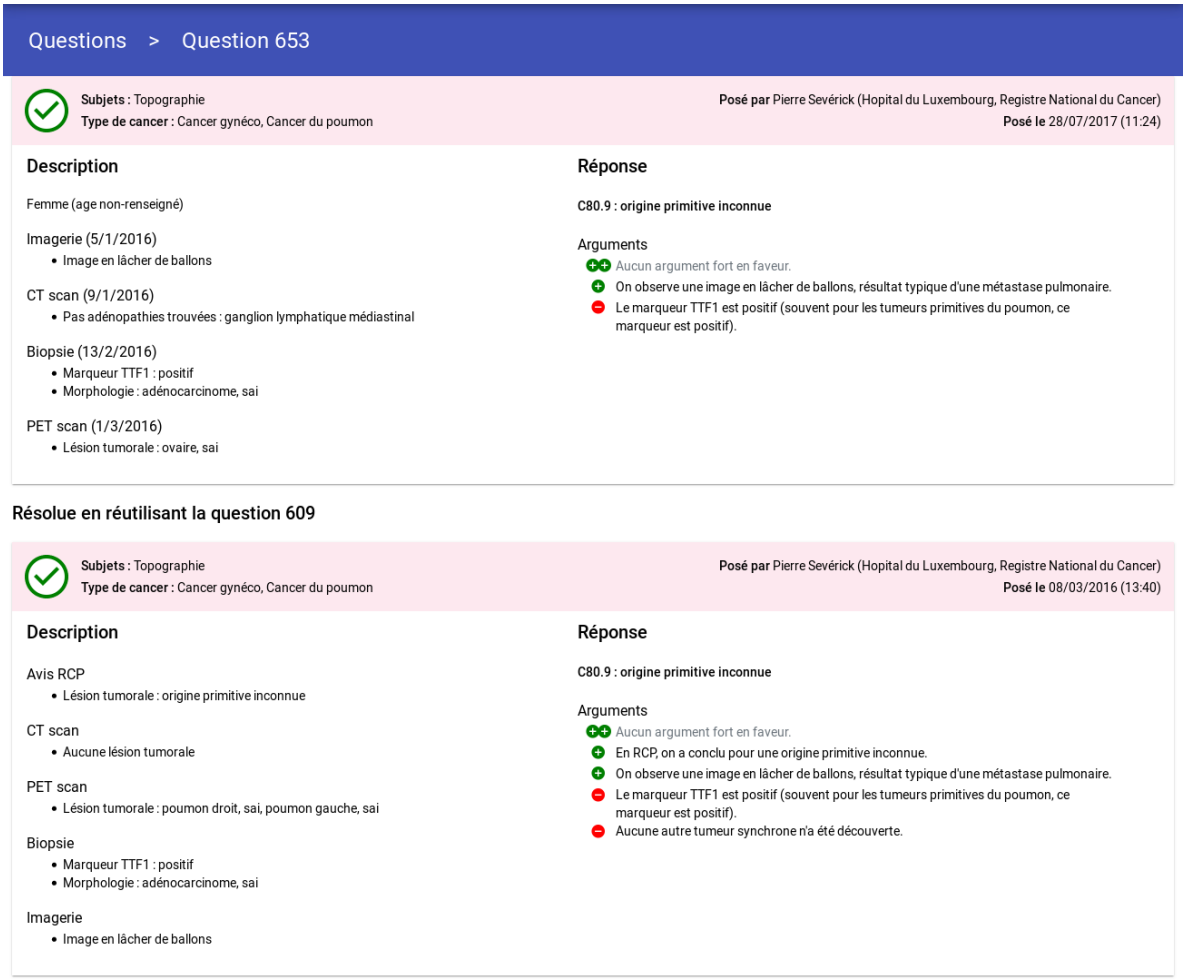


FIGURE *4 – Affichage au sein du prototype de l'exemple décrit dans l'introduction. Le prototype affiche la question, une vue synthétique du dossier patient et, si celle-ci à été fournie, la solution. Pour la solution, la réponse et les arguments sont affichés. Si la réponse a été déduite à partir d'un ancien cas, ce-dernier est affiché pour mieux expliquer la solution et le raisonnement du prototype.*

Le prototype a été testé en interne pour obtenir une première évaluation de son utilisabilité et utilité. Quelques anciennes questions concernant la topographie et les connaissances du domaine associées ont été formalisées. Beaucoup de soin a été donné durant la phase de formalisation afin de rendre les arguments largement applicables. Ensuite de nouvelles questions ont été posées au prototype et les réponses obtenues ont été comparées avec les réponses attendues. Le prototype a fourni une réponse à toutes les questions, néanmoins ces réponses n'étaient pas toujours correctes. Cet écart s'explique essentiellement par le faible nombre de cas dans la base de case (une quinzaine de questions pour les premiers tests,

6. https://angular.io, https://golang.org, https://github.com/gin-gonic/gin
7. https://jena.apache.org/, https://jena.apache.org/documentation/fuseki2/

sachant que ce nombre va augmenter au fur et à mesure de l'utilisation de l'outil) et le mécanisme très simple de réutilisation des cas actuel. En effet, comme les arguments ont été formalisés avec un périmètre d'applicabilité assez large, les réponses fournies peuvent être légèrement incorrectes (par exemple, indiquer une topographie de lobe supérieur de poumon au lieu du lobe inférieur du poumon). Cependant, comme le prototype affiche la solution fournie conjointement avec le cas source utilisé, un encodeur devrait être en mesure de réaliser les adaptations ou corrections nécessaires pour obtenir la réponse attendue. Pour les questions concernant d'autres sujets, le prototype repose entièrement sur les experts de codifications.

## 4 Conclusion

Récemment, il y a un intérêt croissant pour l'application de raisonnement à partir de cas en médecine (Bichindaritz *et al.*, 2015). Cet article propose une approche d'aide aux encodeurs dans l'interprétation des bonnes pratiques de codification médicale. Cette approche a été élaborée suite à des discussions avec les encodeurs et les experts de codification du RNC sur base de cas réels. Une douzaine de cas complexes ont été revus en détail, en plus d'une centaine de cas plus simples. Les questions posées par les encodeurs sont comparées avec d'anciennes questions et sont résolues en réutilisant les arguments en faveur et en défaveur des anciennes solutions. Les résultats présentés sont préliminaires et sont à revoir lors d'une évaluation plus complète, prenant également en compte le retour des encodeurs et des experts de codification.

A ce stade, le processus de raisonnement pour la solution d'une question est partiel. Les arguments utilisés ne représentent qu'une partie d'un processus plus complexe. De même, les cas sources ne sont pas modifiés lors de leur réutilisation. La formalisation du processus de raisonnement, l'intégration des standards de codification et des généralisations des cas réutilisés sont des pistes futures de recherche prometteuses.

Lorsque le prototype aura été validé et amélioré par un usage quotidien, une seconde version, plus indépendante du domaine considéré, sera réalisée. L'objectif à terme est de construire un système générique de raisonnement à partir de cas argumentatif utilisant des standards du web sémantique.

**Remerciements.** Le premier auteur souhaite remercier la Fondation Cancerpour son soutien financier.

## Références

AAMODT A. & PLAZA E. (1994). Case-based reasoning : Foundational issues, methodological variations, and system approaches.

BICHINDARITZ I., MARLING C. & MONTANI S. (2015). Case-based Reasoning in the Health Sciences. In *Workshop Proceedings of ICCBR*.

BRICKLEY D. & GUHA R. V. (2014). RDF Schema 1.1, `https://www.w3.org/TR/rdf-schema/`, W3C recommendation, last consultation : March 2017.

BUNKE H. & MESSMER B. T. (1993). Similarity measures for structured representations. In *European Workshop on Case-Based Reasoning*, p. 106–118 : Springer.

EUROPEAN NETWORK OF CANCER REGISTRIES AND TYCZYNSKI, JERZY E AND DÉMARET, D AND PARKIN, D MAXWELL (2003). *Standards and guidelines for cancer registration in Europe : the ENCR recommendations*. International Agency for Research on Cancer.

MAXIMINI K., MAXIMINI R. & BERGMANN R. (2003). *An investigation of generalized cases*, In *Case-Based Reasoning Research and Development*, p. 261–275. Springer.

RICHTER M. M. & WEBER R. O. (2013). *Case-based reasoning : a textbook*. Springer Science & Business Media.

SCHNELL M., COUFFIGNAL S., LIEBER J., SALEH S. & JAY N. (2017). Case-Based Interpretation of Best Medical Coding Practices — Application to Data Collection for Cancer Registries. In *Conference Proceedings of ICCBR*.