Utilisation de métatermes pour la recherche d'information dans les dossiers médicaux

Pereira Suzanne[1], Philippe Massari[2], Michel Joubert[3] et Stefan Darmoni [1]

[1] Groupe GCSIS, Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes (LITIS), Université de Rouen, France
[2] Unité Informatique médicale, CHU de Rouen, Rouen, France
[3] Laboratoire d'Enseignement et de Recherche sur le Traitement de l'Information Médicale (LERTIM), Faculté de médecine de Marseille, Marseille, France

Abstract

Objectives: *In order to create views of the medical records by medical specialities or by medical nature of the procedures, we investigate to use a terminological indexing of the stay's coding (ICD10 for the diseases and CCAM (the French CPT (common procedure terminology)) for the procedures)*

Methods: *To reach this objective, we assign automatically and manually super-concepts for each terms of the classifications (ICD10 and CCAM). The super-concepts were created above the MeSH thesaurus by the CISMeF Team. We then compare the automatic and manual assignations.*

Results: *These comparisons provide a precision of 21 to 60% and a recall of 28 to 59% depending on the methods and on the nomenclatures chosen.*

Conclusions: *In conclusion, we show that the assignations of super-concepts could be done manually and automatically for these two nomenclatures. In a future work we will try to demonstrate that they can be applied to display a more relevant presentation of the medical records.*

Keywords

Medical Subject Headings; medical records, problem-oriented; vocabulary, controlled; Healthcare Common Procedure Coding System; International Classification of Diseases.

1 Introduction

Les dossiers médicaux contiennent les éléments définis par la législation en vigueur. En France, la loi du 4 mars 2002 et du décret du 27 avril 2002 définissent le contenu du dossier médical.

Les dossiers médicaux informatisés (DMI) sont plus ou moins structurés en fonction des applications utilisées, leur architecture est définie en Europe par la norme HISA [1]. Les niveaux patient, séjour ou prise en charge (CONTACT d'HISA) et actes sont pris en compte dans la quasi-totalité des applications hospitalières en France. La notion d'épisode (CASE d'HISA), entité de regroupement est moins souvent présente et n'est utilisée que dans des cas particuliers (par exemple les séances de chimiothérapie).

La présentation des éléments des DMI a fait l'objet de longue date de nombreuses publications, [2,3,4,5,6,7]. Trois typologies y sont retrouvées :

1. Le classement chronologique (habituellement inverse) est le plus naturel et le plus classique.
2. Le dossier orienté problème, nécessite de relier chacun des éléments du dossier à une des pathologies du patient (théorie ancienne proposée par Weed en 1968 [2]). Un séjour est associé à un ou plusieurs problèmes. Si N séjours sont associés au même problème alors ces séjours sont regroupés en un épisode (HISA).
3. Les dossiers classés en fonction de la nature médicale des éléments, comptes-rendus de séjours, actes et comptes-rendus de radiologie, d'anatomie pathologique…

La présentation chronologique est parfaitement adaptée à la consultation des dossiers peu volumineux. La recherche d'éléments dans des DMI de patients pris en charge de longue date et à de nombreuses reprises (dans notre expérience rouennaise 10.2% des dossiers contiennent plus de 50 séjours et de 100 actes médicaux) est laborieuse voir impossible.

Dans l'idéal, les deux autres modalités d'accès aux éléments de DMI devraient y être associées. Le « dossier orienté problème » nécessite la saisie des pathologies, elle est possible au sein d'un service ou pour un mono utilisateur du DMI, mais a toujours posé un problème au sein d'établissements importants [8].

La classification par la nature médicale des éléments est une méthode reproduisant le classement habituel des dossiers papiers, nécessitant des traitements informatiques spécifiques dépendant de l'organisation de l'établissement.

Ces trois modes sont complémentaires. Des vues du dossier correspondant aux modes 2 et 3 semblent réalisables en se fondant sur des extensions terminologiques.

L'objectif de ce travail est d'étudier la possibilité d'utiliser une indexation terminologique des codages des séjours en CIM10, des actes en CCAM et des entités (unités fonctionnelles) prenant en charge les patients ou pratiquant les actes. Et ceci afin de construire des vues de dossiers présentant des éléments des DMI en fonction des pathologies ou de la nature médicale des éléments. Nous étudierons dans ce travail des approches manuelles et automatiques.

Ce travail se fonde sur une expérience de plus de 10 ans de l'équipe CISMeF autour du thésaurus MeSH notamment l'adjonction de métatermes (que nous décrivons plus bas). Dans ce travail, nous appliquerons l'idée des super-concepts sur deux autres terminologies : la CIM10 et la CCAM.

2 Etat de l'art : les données terminologiques à notre disposition

2.1 MeSH

Le thésaurus MeSH (Medical Subject Headings, NLM (National Library of Medicine),1960) est une terminologie avec un vocabulaire contrôlé qui a été conçue pour indexer les articles scientifiques dans la base de données bibliographique MEDLINE [9]. Les relations entre concepts les plus fréquentes sont des relations de spécialisation – généralisation, tout-partie (méronymie). Dans sa version 2006, il comporte 22.995 mots-clés et 61.000 synonymes (dont 7000 ajoutés par l'équipe CISMeF) répartis en 15 arborescences thématiques auxquelles correspondent un code spécifique (exemple : 'maladie' C) qui peuvent être associés à l'un des 83 qualificatifs (exemple : 'diagnostic', 'prévention & contrôle') afin de préciser leur sens. Les mots-clés possèdent un identifiant

unique et un code dépendant de leur place dans l'arborescence (exemple : 'amyloïdose' : D000686 et C18.452.090).

Le thésaurus MeSH dans sa structure d'origine, ne permet pas d'obtenir de vision globale d'une spécialité médicale. L'équipe CISMeF a ainsi adapté le MeSH pour répondre à cette problématique et afin d'adapter le MeSH au domaine plus large des ressources de santé sur l'Internet [10]. L'équipe a developpé plusieurs améliorations au MeSH depuis 2000. En plus des mots-clefs MeSH et des qualificatifs, le concept de métaterme (MT) (n=135) et les types de ressources (RT) (n=274) ont été ajoutés.

Un MT est généralement une spécialité médicale ou une science biologique qui a des liens sémantiques avec 1 ou plusieurs termes MeSH, qualificatif, et RT (ex : cardiologie, bacteriologie).

Les métatermes ont été sélectionnés manuellement par le conservateur des bibliothèques de l'équipe CISMeF (B. Thirion) en s'appuyant également sur l'expertise de spécialistes CHU de Rouen. Pour chaque métaterme, différents liens sémantiques ont été créés. Ces métatermes peuvent être considérés dans la terminologie CISMeF comme des super-concepts qui permettent une vision plus globale concernant une spécialité en offrant un niveau supplémentaire d'abstraction. Il y a de 0 à n relations entre les termes CISMeF et les métatermes CISMeF.

L'idée des métatermes a été créée afin d'optimiser la recherche d'information dans CISMEF en contournant la nature relativement restreinte des termes MeSH correspondant à une spécialité et permettre de créer à la demande des filtres CISMeF par spécialités. Les métatermes permettent en effet de connaître l'ensemble des termes MeSH qui sont repartis dans plusieurs arborescences mais qui concernent une même spécialité biologique ou médicale. Par exemple, dans les requêtes 'recommandation en cardiologie »' et 'base de données en psychiatrie', les termes 'psychiatrie' et 'cardiologie' ne sont considérés que comme des mots-clefs MeSH, ce qui donne très peu ou peu de réponse dans CISMeF. Introduire cardiologie et virologie comme des métatermes est une technique efficace pour retrouver plus de résultats parce qu'au lieu d'exploité une seule arborescence MeSH (ex : psychiatrie comme un mot-clef MeSH), le fait d'utiliser les métatermes entraîne une expansion automatique des requêtes en exploitant d'autres arborescences MeSH ou CISMeF grâce aux liens sémantiques. (exemple : 'hôpital psychiatrique' comme un mot clés MeSH ou 'hôpital psychiatrique' comme type de ressource seront exploités dans le cas d'une requête sur la 'psychiatrie').

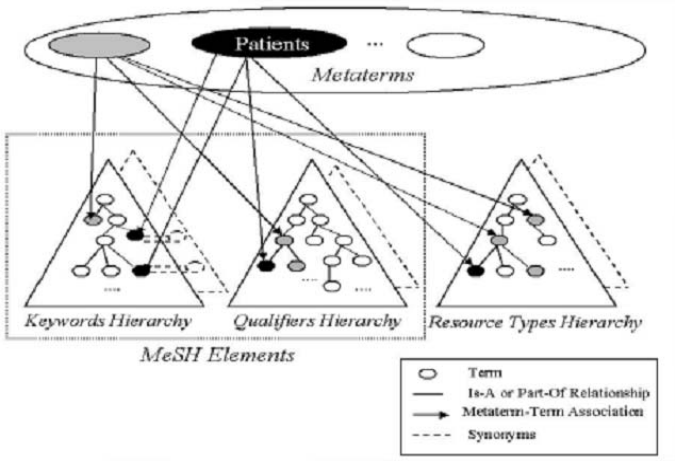

Figure 1 : Les liens sémantiques entre les métatermes CISMeF et les termes MeSH

2.2 CIM10

Les diagnostics dans les dossiers médicaux sont codés à l'aide de la CIM10 (la Classification Statistique Internationale des Maladies et des problèmes de santé connexes 10e version [11]) qui est ordonnée en une hiérarchie à 5 niveaux maximum partitionnés en 21 chapitres couvrant l'éventail complet des états morbides classés par appareil fonctionnel et associé à une lettre (exemple : E : 'Maladies endocriniennes, nutritionnelles et métaboliques'). Les chapitres sont divisés en groupes, eux-même divisés en sous-groupes composés de catégories à 3 et de sous catégories à 4 caractères, englobant le contenu des termes CIM-10. La CIM10 inclus plus de 18.000 codes alphanumériques et environ 50.000 termes. Par ailleurs, des extensions de codes de la CIM-10 ont été créées pour le PMSI par le PERNNS (Pôle d'expertise et de référence national des nomenclatures de santé) et l'ATIH (Agence Technique de l'Informatisation sur l'Hospitalisation) pour apporter plus de précision à certains codes et améliorer le classement en GHM (ex : M45.+4, S82.00, E10.8A). Les diagnostics pour les tumeurs peuvent être codés en CIMO (la CIM oncologie) (ex : M8020/3 carcinome indifférencié SAI).

2.3 CCAM

La CCAM (Classification Commune des Actes Médicaux [12]) est le nouveau référentiel des actes médicaux qui remplace, pour les médecins, la Nomenclature générale des actes professionnels (NGAP) en secteur libéral, et le Catalogue des actes médicaux (CDAM) en secteur hospitalier. Élaborée par la CNAMTS et l'ATIH, en étroite collaboration avec les sociétés savantes, la CCAM constitue une liste d'actes codés, commune aux secteurs public et privé.

Elle est destinée à décrire plus précisément chaque acte, à servir de base à la tarification en secteur libéral (cabinets et cliniques) et à l'allocation des ressources aux établissements publics dans le cadre de la tarification à l'activité (T2A).

A chaque libellé de la CCAM correspond un code à 7 caractères alphanumériques : les 4 premiers sont signifiants (topographie, action, voie d'abord et/ou technique), les 3 derniers constituent un compteur séquentiel.

AA | AA | NNN
topographie action Voie d'abord et/ou technique compteur

Exemple : GEFA012 : 'résection anastomose' // 'de trachée' // s'ans abaissement du larynx' // 'par cervicotomie' (Action topographie technique voie d'abord). GE : 'trachée' ; F: 'résection' ; A: 'abord ouvert'

3 Matériel et méthodes

3.1 Ajout des super-concepts aux nomenclatures CIM10 et CCAM

3.1.1 Ajout des super-concepts à la CIM10

Les métatermes ont été définis manuellement par un expert (P. Massari) en utilisant la hiérarchie de la nomenclature. Pour chaque sous-chapitre de dernier niveau, il a été défini un ou plusieurs métatermes lorsqu'ils s'appliquaient aux codes sous-jacents. Dans un certain nombre de cas des métatermes ont été définis au niveau des codes, soit en complément, soit quand aucun n'était adapté à tous les codes d'un chapitre.

Nous avons automatiquement associé des métatermes aux codes CIM10. Nous avons pour cela utilisé la table de transcodage CIM10/MeSH extraite du métathésaurus de

l'UMLS (Unified Medical Language System [13]). Cette table permet de retrouver à partir d'un code CIM10 le ou les mots clefs MeSH supposés équivalents au terme CIM10. Cette méthode est limitée puisque tous les codes CIM10 n'ont pas d'équivalent en MeSH. Seul 8.9% des codes CIM10 sont transcodables. Nous avons déterminé la liste des termes MeSH supposés équivalents aux codes CIM10 pour les codes CIM10 transcodables. Puis nous avons obtenu la liste des métatermes reliés à ces termes MeSH.

3.1.2 Ajout des super-concepts au CCAM

La CCAM est classée par grands appareils et non par spécialités ce qui ne permet pas d'emblée de définir un métaterme pour les codes. L'utilisation du modèle GALEN [12] donne une signification au code lui-même par les quatre lettres qu'il contient (voir chapitre 2.3), les deux premières correspondent à une région anatomique, la troisième à l'action, la quatrième à la voie d'abord. L'utilisation des deux premières lettres du code et de la dernière a permis à un expert de définir manuellement les mots-clefs MeSH, et ainsi développé un transcodage CCAM/MeSH, et les métatermes au niveau des groupes de codes ainsi constitués. Dans un deuxième temps, les mots-clefs et les métatermes ont été revus pour chaque code par l'expert.

A la suite de cela, nous avons automatiquement associé des métatermes aux libellés CCAM et ceci de deux façons :

• Tout d'abord, nous avons utilisé l'algorithme du sac de mot [14,15] sur les libellés CCAM. Cet algorithme est utilisé dans le catalogue CISMeF pour la recherche d'information, pour retranscrire les requêtes de l'utilisateur qui sont faites en langage naturel en termes MeSH et ainsi permettre au système de proposer à l'utilisateur, des documents correspondants à la requête (indexés avec ces termes MeSH). Pour le CCAM, cet algorithme procède comme suit :

Chaque terme CCAM est segmenté en mots, les mots vides (les termes considérés non pertinents exemple : 'le', 'la') sont supprimés, le terme est alors considéré comme un ensemble de mots (l'ordre n'y est plus pertinent). Le principe de base de l'appariement est de proposer les termes MeSH qui contiennent le maximum de mots du terme CCAM sans tenir compte de leur position. Plusieurs termes cibles peuvent être nécessaires pour couvrir les différents mots d'un terme. Tour à tour, le terme MeSH couvrant le maximum des mots restants de la requête est sélectionné. Cet algorithme utilise la phonémisation ce qui permet d'étendre l'appariement entre les mots et les termes MeSH aux formes fléchies des mots.

Nous avons ainsi extrait les mots-clefs MeSH contenus dans chaque libellé CCAM. Ces mots-clefs MeSH sont reliés aux métatermes par des liens sémantiques (voir fig1). Nous avons ainsi pu déterminer les métatermes associés à chaque liste de termes MeSH pour chaque libellé CCAM. Ces métatermes avant dédoublonnement pouvant être nombreux (15 alors que l'expert a associé en moyenne 1.18 métatermes par libellé CCAM), et plusieurs mots-clefs MeSH d'une même liste pouvant être associé au même métaterme, nous avons décidé arbitrairement de calculer la fréquence pour chaque métatermes obtenus et de ne prendre que les deux métatermes les plus fréquents pour chaque liste de métatermes.

• Dans une deuxième étude, nous avons utilisé les mots-clefs MeSH associés manuellement aux libellés CCAM par l'expert pour retrouver les métatermes reliés. De la même façon nous n'avons pris en compte que les deux métatermes les plus fréquents.

3.2 Comparaisons des assignations de métatermes automatiques et manuelles

Nous avons comparé les assignations de métatermes faites automatiquement et manuellement par notre expert. Le "gold standard" pris a été le codage manuel.

Nous avons ainsi calculé le nombre de vrai positifs, faux positifs, faux négatifs et la précision (moyenne pour tous les libellés) et le rappel (moyenne pour tous les libellés) qui sont les mesures de référence dans le domaine des sciences de l'information (voir tableau 1).

Tableau 1 : Tableau des différentes mesures utilisée pour les comparaisons des assignations de métatermes automatiques et manuelles

Mesures	Définition et calcul
Vrai positifs	Nombre de métatermes correctement retournés automatiquement et pertinents (assignés par l'expert)
Faux négatifs	Nombre de métatermes non retournés automatiquement mais pertinents (assignés par l'expert)
Faux positifs	Nombre de métatermes retournés automatiquement mais non pertinents (non assignés par l'expert)
Précision	Proportion de métatermes pertinents dans les métatermes retournés automatiquement = Vrai positifs / Nombre de métatermes retournés (nb de métatermes assignés automatiquement)
Rappel	Proportion de métatermes retournés dans les métatermes pertinents (retournés par l'expert) = Vrai positifs / Nombre de métatermes pertinents (assignés par l'expert)

Nous avons pris en compte les métatermes identiques et les associations de métatermes, c'est-à-dire lorsqu'un métaterme est lui-même issu de la combinaison de deux autres métatermes (exemple : 'chirurgie' + 'neurologie' = 'neurochirurgie' ; 'chirurgie orthopédique' = 'orthopédie' + 'chirurgie'). Ainsi si l'expert a assigné le métaterme neurochirurgie à un code CCAM, et que automatiquement nous avons assigné les métatermes chirurgie et neurologie, alors lors de l'évaluation ces deux assignations seront considérées comme équivalentes et comme un vrai positif.

Nous avons pu observer que les métatermes distincts utilisés par l'expert était différents de ceux assignés automatiquement par nos différentes méthodes. … métatermes ont été utilisés par l'expert, les autres ayant été considérés comme peu utile pour la recherche de documents dans le dossier patient. Nous avons ainsi, dans une deuxième étude, pour nos différentes mesures, considéré seulement les métatermes utilisés par l'expert parmis tous les métatermes extraits automatiquement.

4 Résultats

4.1 Assignations des super-concepts au termes CIM10 et CCAM

4.1.1 Pour la CIM10

Dans un premier temps, les métatermes ont été assigné manuellement par l'expert aux groupes de codes CIM10 (exemple : A10). Il a ainsi été crée 2.311 paires groupe de codes CIM10 / métaterme à partir de 1.590 groupe de codes CIM10. L'expert a assigné de 1 à 4 métatermes pour chaque code avec une moyenne de 1,45 (+/- 0,63) métatermes par groupe de codes.

Dans un deuxième temps, les métatermes ont été définis au niveau des codes CIM10 de dernier niveau soit en complément, soit quand aucun n'était adapté à tous les codes d'un chapitre. 1.219 paires code CIM10 de dernier niveau/métaterme ont été crées pour 1.121 codes CIM10 de dernier niveau. 1 à 3 métatermes ont été assignés pour chaque code avec une moyenne de 1,09 (+/- 0,32) metatermes par code.

En appliquant les assignations de métatermes définis pour chaque groupe de codes à tout le groupe, on trouve au total 13.650 paires code CIM10 / métaterme pour 10.505 codes CIM10 avec l'assignation de 1 à 5 métatermes pour chaque code (moyenne de 1,52 (+/- 0,67) métatermes par code).

4.1.2 Pour la CCAM

Au niveau de la CCAM, 8.698 paires libellé CCAM / métaterme ont été produits manuellement par l'expert à partir des 7.389 libéllés de la CCAM. 0 à 4 avec une moyenne de 1,18 (+/- 0,50) métatermes ont été définis pour chaque libellé CCAM. L'expert n'a pas pu associé de métaterme pour 126 libellés.

4.2 Comparaisons des assignations de métatermes automatiques et manuelles

4.2.1 Utilisation du transcodage MeSH pour la CIM10

1.542 codes CIM10 étaient transcodables en MeSH parmi les 10.505 codes du départ (soit 14,68%). Nous avons ainsi défini automatiquement 2.250 paires code CIM10 / métaterme. 0 à 7 métatermes ont été assignés pour chaque code avec une moyenne de 1.46 (+/-1,32) métatermes par code. Nous n'avons pas pu trouvé de métaterme reliés aux termes MeSH pour 109 codes CIM10. Comparé au codage manuel fait par l'expert, nous avons trouvé une précision de 60% et un rappel de 59% (voir tableau 2).

Tableau 2 : Comparaison entre les métatermes assignés manuellement et automatiquement (à l'aide du transcodage CIM10/MeSH)

Mesures	En considérant tous les métatermes	En ne considérant que les métatermes pertinents
Vrai positifs	1.348	1.348
Faux négatifs	948	948
Faux positifs	902	819
Précision	0,49	0,60
Rappel	0,58	0,59

4.2.2 Utilisation de l'algorithme du sac de mots pour la CCAM

Grâce à l'algorithme du sac de mots, il a été trouvé automatiquement 13946 paires libellé CCAM / métaterme à partir de 7.389 libéllés CCAM. 0 à 11 métatermes ont été définis pour chaque libellé avec une moyenne de 1,89 (+/- 1,59) metatermes par libellé. Pour 1.150 libellés CCAM nous n'avons pas pu extraire de métaterme. Comparé aux assignations faites par l'expert, il a été trouvé une précision de 21% et un rappel de 28% (voir tableau 3).

Tableau 3 : Comparaison entre les métatermes assignés par l'expert et automatiquement(avec l'algorithme du sac de mots)

Mesures	En considérant tous les métatermes	En ne considérant que les métatermes pertinents
Vrai positifs	2.439	2.439
Faux négatifs	6.307	6.307
Faux positifs	11.467	6.303
Précision	0,17	0,21
Rappel	0,27	0,28

4.2.3 Utilisation des mots-clefs MeSH originellement assignés pour chaque libéllé CCAM pour la CCAM

A partir des mots-clés assignés manuellement par l'expert il a été définis automatiquement 1 à 10 métatermes (avec une moyenne de 2,08 (+/- 1,39) metatermes par libéllé) pour chaque libéllé soit 15.400 paires libéllé CCAM / métaterme à partir de 7.389 libéllés CCAM. La comparaison de ces assignations automatiques avec les assignations manuelles fournit une précision de 29% et un rappel de 38% (voir tableau 4).

Tableau 4 : Comparaison entre les métatermes assignés par l'expert et automatiquement (à l'aide des mots-clefs MeSH assigné manuellement aux libéllés CCAM)

Mesures	En considérant tous les métatermes	En ne considérant que les métatermes pertinents
Vrai positifs	3.239	3.239
Faux négatifs	5.459	5.459
Faux positifs	12.136	6.715
Précision	0,14	0,29
Rappel	0,34	0,38

5 Discussion

L'utilisation des métatermes CISMeF afin de présenter le contenu de DMI en fonction de regroupements pathologiques ou de la nature médicale des éléments nécessite l'indexation terminologique des codes CIM10 et CCAM.

Les objectifs de ce travail étaient d'étudier la possibilité de générer automatiquement des métatermes liés aux codes CIM10 et CCAM, en comparant l'indexation automatisée à celle réalisée par un "expert" à la main, et de faire une première évaluation des potentialités de ce type d'indexation à réaliser des fonctions de sélection et de présentation des informations pertinentes des DMI.

5.1 Discussion méthodologique

Les méthodes d'indexation utilisées étaient de 3 types :

- méthode manuelle pour la CIM10 et la CCAM.
- méthode automatisée à partir des transcodages validés par des experts MeSH/CIM10 (extrait de l'UMLS) et MeSH/CCAM (réalisé précédemment par notre expert).
- méthode automatisée à partir des libellés grâce à l'algorithme du sac de mot pour la CCAM.

Les comparaisons des assignations de métatermes manuelles et automatiques, tant pour la CIM10 que pour la CCAM, montrent des différences. Ces comparaisons ont montré une précision de 21 à 60% et un rappel de 28 à 59% selon les méthodes et les nomenclatures utilisées.

De manière générale, les résultats peuvent s'expliquer par le fait que l'expert a assigné des métatermes dans un objectif de recherche dans un dossier médical fondée sur la pratique médicale, alors que les méthodes automatiques se fondent sur les relations métaterme CISMeF/ mots clés MeSH qui avaient été originellement utilisées dans un objectif de recherche documentaire dans CISMeF. De ce fait, les assignations des métatermes peuvent être, pour la CIM10 et la CCAM, très différentes selon qu'elles soient faites manuellement ou en passant par des méthodes automatiques.

Pour pallier en partie ce problème, nous avons, dans une deuxième étude, pris en

compte dans nos évaluations que les métatermes appartenant à la liste globale des métatermes distincts utilisée par l'expert. Les résultats des indexations automatiques sont alors meilleurs.

Les nombres de métatermes extraits manuellement et automatiquement sont aussi différents pour chaque libellé CCAM ou CIM10. Les méthodes manuelles peuvent générer plus de métatermes que nos méthodes automatiques et vis et versa. Les précisions et rappel sont alors pour chaque libellé souvent différentes de 1.

De plus, dans cette étude, nous avons choisi comme gold standard l'indexation manuelle des libellés CCAM et CIM10. Les métatermes utilisés sont proches des spécialités médicales, dont les contours ne sont pas toujours très bien définis et dépendent de pratiques "locales". Une grande variabilité inter-expert dans l'assignation de ces métatermes est dans ce cadre tout à fait vraisemblable.

Pour chaque méthode, nous avons pu identifier d'autres raisons plus spécifiques qui expliquent ces résultats.

5.2 Indexation à partir des transcodages

C'est cette méthode qui donne les meilleurs résultats avec des taux de précision et de rappel de l'ordre de 0,5 et 0,6 pour la CIM10 et de 0,3 et 0,4 pour la CCAM. En effet, elle utilise les tables de transcodages validées manuellement qui permettent de définir des termes MeSH supposés équivalents ou englobant les termes CIM10 et CCAM.

Il a été déterminé pour cette méthode, un nombre de **faux positifs** (voir tableau 1) importants, ils peuvent être expliqués par le fait que certains mots clés sont retrouvés dans plusieurs arborescences MeSH, liées sémantiquement à plusieurs métatermes. Certains de ces métatermes peuvent ne pas s'appliquer pour certains actes ou maladies très spécifiques. D'autres peuvent être considérés par l'expert comme peu pertinent.

Exemple : Pour indexer le code CCAM 'GDNE001 - Coagulation d'un œdème du larynx, par laryngoscopie directe avec laser', l'expert a entre autre utilisé le mot clé 'larynx' sémantiquement lié au métaterme 'oto-rhino-laryngologie', et appartenant à l'arborescence MeSH 'appareil respiratoire' lié à 'pneumologie'. L'expert n'a retenu qu"oto-rhino-laryngologie', alors que l'indexation automatique donne en plus le métaterme 'pneumologie'.

Mais aussi et particulièrement pour la CIM10 lorsque les mots clés génèrent un métaterme correspondant à une spécialité d'organe, alors que s'agissant d'une pathologie exclusivement infantile, l'expert n'a retenu que le métaterme 'pédiatrie'.

Dans ce dernier cas outre un faux positif, l'indexation automatique entraîne un faux négatif. Le nombre de **faux négatifs** aussi bien pour la CIM10 que pour la CCAM est important, ils semblent correspondre dans bon nombre de cas à ce type de mécanisme, l'expert a utilisé un métaterme, alors que l'algorithme en ramène un ou plusieurs qui ne correspondent pas. C'est le cas par exemple lorsque l'expert choisi d'englober les différents concepts inclus dans les libellés dans un métaterme beaucoup plus général.

Exemple : pour le terme CCAM QAEA001 'Transplantation de moins de 50 greffons de cuir chevelu'. Automatiquement nous avons trouvé les métatermes 'thérapeutique', 'transplantation' et 'chirurgie'. Alors que l'expert a trouvé 'chirurgie plastique reconstructrice et esthétique' qui englobe tous ces métatermes.

Ce nombre de faux négatifs et faux positifs peuvent aussi être expliqué par le fait que le transcodage CIM10/MeSH peut produire des termes MeSH plus précis ou plus globaux que ceux utilisés originellement dans les libellés CIM10. Ce qui peut entraîner la reconnaissance automatique de métatermes en plus ou en moins par rapport aux assignations manuelles.

Plus généralement, seul 8,9% de la CIM10 est transcodable en MeSH, il n'est donc pas

possible de générer automatiquement les métatermes associés à tous les termes de la CIM10 avec cette technique. Néanmoins, parmi les 1000 codes CIM10 les plus codés au CHU de Rouen, 53,5% sont transcodables en MeSH et appartiennent à notre table. Ces 1000 codes couvrent 82,03% des lettres de sortie.

5.3 L'indexation par l'algorithme du sac de mots

Cette méthode purement lexicale en pratique est la plus intéressante, car elle ne nécessite aucune indexation manuelle. Par contre, elle montre de moins bons résultats.

Dans son principe elle ramène tous les termes MeSH à partir du libellé du code CCAM, ces mots clés permettent de définir les métatermes selon les mêmes principes que la méthode précédente.

Une partie des faux positifs et des faux négatifs dans les métatermes générés répond aux mêmes mécanismes que la méthode d'indexation à partir des transcodages.

Le nombre de faux positifs est majoré par le fait qu'à l'inverse de la méthode précédente les mots clés MeSH extraits à partir des libellés ne sont pas validés, l'expert n'a pu choisir de ne pas prendre en compte certains termes, la méthode automatisée prend en compte tous les termes MeSH qu'elle arrive à trouver.

Le nombre important de faux négatifs peut en partie être expliqué par une faible adéquation sémantique entre la terminologie CCAM et le MeSH, et par le fait que l'algorithme du sac de mot ait été développé pour une indexation documentaire et non dans un but de classification d'actes techniques.

Exemple : pour terme CCAM 'ABFA007 « Exérèse d'une fistule dermique avec prolongement intradural occipital' l'algorithme du sac de mots n'a pas permis d'extraire le métaterme 'neurochirurgie', il a extrait les métatermes 'rhumatologie' et 'dermatologie'.

Le choix de ne prendre que les deux métatermes les plus fréquents pour les assignations automatiques peut également être une explication du nombre de faux négatifs. Certains métatermes ne sont pas pris en compte parce que les termes MeSH auxquels ils sont rattachés étaient lexicalement moins présents dans le libellé ou au niveau des liens entre les mots clés MeSH et les métatermes. La fréquence n'est peut-être pas le bon critère de sélection des métatermes, une pondération des métatermes ou des mot clés pourraient être plus performante.

5.4 Autres méthodes

Dans notre évaluation, certains termes sont considérés comme des faux positifs alors que ce sont de vrais positifs. Ces termes qui ont été reconnus automatiquement mais oubliés dans l'indexation manuelle pourraient être rajouté à l'indexation manuelle. Nous pourrions donc, dans une future étude, faire une validation secondaire qui marquerait ce type de métaterme et que nous pourrions ensuite dans une deuxième série de comparaison entre les assignations manuelles et automatiques être rajoutés à l'indexation manuelle.

Exemple : Pour indexer le code CIM10 'Shigellose' les mot clés 'dysenterie bacillaire' et 'shigella' sont utilisés.Le mot clé 'dysenterie bacillaire' appartient aux arborescences : 'infections bactériennes et mycoses' (liée sémantiquement au métaterme 'maladies infectieuses') et 'maladie de l'appareil digestif' (liée à gastroentérologie), le mot clé 'shigella' est dans l'arborescence bactéries (liée à bactériologie). L'expert n'a pris en compte que les métatermes 'maladies infectieuses' et 'bactériologie' et a oublié le métaterme 'gastroentérologie'.

Dans une future étude, nous pourrions aussi étudier la répartition des vrais positifs par métatermes ou appliquer l'algorithme du sac de mot sur les libéllés de la CIM10, ce qui donnerait peut–être des meilleurs résultats puisque l'adéquation terminologique entre la CIM10 et le MeSH est plus grand que celle entre le MeSH et la CCAM, le MeSH ayant été

crée à la base à partir de la CIM.

Le fait de s'intéresser à la nomenclature CCAM était très intéressante mais nous étions forcé de constater que cela ne fait que depuis un an que la CCAM a été mise en place dans les hôpitaux avec l'arrivée de la T2A [16]. Tous les anciens dossiers étaient codés en CDAM. Plus récent que la NGAP, le CDAM a vu le jour lors de la mise en place du Programme de médicalisation des systèmes d'information (PMSI), en 1985. La réglementation contraint les établissements de soins et les professionnels à utiliser simultanément ces deux nomenclatures conçues pour des objectifs différents. Pour pallier à ce problème, dans un objectif opérationnel de naviguer plus facilement dans les dossiers de patients complexes, l'expert a aussi assigné manuellement des métatermes à partir de la CDAM.

5.5 Applications et perspectives

La problématique d'accès à la bonne information, que l'on retrouve déjà dans les logiciels métiers qui se contentent le plus souvent de faire de l'entassement chronologique des données, risque d'être majorée dans le futur Dossier Médical Personnel. Souhaiter améliorer la démarche de soins et le suivi des patients avec ce e-DMP, impose que chaque médecin puisse disposer d'outils performants de sélection et de présentation des informations pertinentes [17].

L'indexation dans une terminologie commune des éléments des dossiers médicaux informatisés, peut s'appuyer sur l'indexation des nomenclatures utilisées pour coder les actes et les diagnostics (CCAM, CDAM, CIM10 …).

L'utilisation de métatermes, nous semble pouvoir être la base de la réalisation de fonctions de sélection et de présentation des informations pertinentes des dossiers médicaux codés, ceci restant néanmoins à démontrer. L'accès contextuel à des ressources documentaires, à partir de dossiers médicaux pourrait également reposer sur ce type d'indexation.

6 Conclusion

Ce travail montre toute la difficulté de l'indexation des nomenclatures médicales, une indexation initiale par la méthode du sac de mot, avec un algorithme adapté à la classification d'éléments de dossiers, pourrait être une aide à l'indexation des experts.

Remerciements

Les auteurs remercient la société Vidal qui finance le travail de thèse de Suzanne Pereira dans lequel s'inscrit ce travail.

Références

[1] Groupement pour la Modernisation du Système d'Information Hospitalier (GMSIH). Les annuaires normes et standard version 1.0. *IN11NSS10*. 2003

[2] Weed LL. Medical records that guide and teach. *N. Eng J Med* 1968; 278: 593-600 et 652-7

[3] Bayegan E., Nytrø, and Grimsmo, A. An Ontologies for Knowledge Representation in the Computer-Based Patient Record System. *ICTAI2002*.

[4] ANAES. la tenue du dossier médical en médecine générale : état des lieux et recommandations. *Recommandations de l'ANAES*. sept 1996

[5] Falcoff H. Dossier médical en médecine générale. *La revue du Praticien Médecine Générale*. Dec 1997; 404:71-8.
[6] Falcoff H. Le dossier orienté problème existe, je l'ai rencontré. *L'informatisation du Cabinet Médical du Futur*. 1999.
[7] Bainbridge M., Salmon P., Rappaport A., Hayes G., Williams J., Teasdale S. The Problem Oriented Medical Record - just a little more structure to help the world go round?. *The Primary Health Care Specialist Group of the British Computer Society*. 1996.
[8] Interopérabilité des dossiers de santé informatisés et normalisation *Etude pour le Conseil Supérieur des Systèmes d'Informations de Santé*. juillet 1999.
[9] Dailland F. Tout Savoir sur le MeSH…ou presque….MeSH (Medical Subject Headings) et FmeSH (version française). Mai 2005.
[10] Douyere M, Soualmia LF, Neveol A, Rogozan A, Dahamna B, Leroy JP,Thirion B, Darmoni SJ. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality –controlled gateway. *Health Info Libr J*. 2004 Dec;21(4):253-61
[11] CIM10 organisation modiale de la santé (OMS) Genève 1993.
[12] Rodrigues J.M., Trombet-Paviot B., Martin C. et Vercherin P. Représentation du standard européen de terminologie EN1828 et de Galen CCAM avec l'éditeur d'ontologie Protégé : vers un système terminologique de troisième génération pour les interventions chirurgicales. *JFIM2005*. 2005.
[13] UMLS fact sheet : http://www.nlm.nih.gov/pubs/factsheets/umls.html
[14] Zweigenbaum P., Darmoni SJ., & Grabar N. The contribution of morphological knowledge to French MeSH mapping for information retrieval. *Journal of the American Medical Informatics Association*. 8(suppl):796-800, 2001.
[15] Soualmia L., Dahamna B., Thirion B., Darmoni S. Some Stratégies for Health Information Retrieval. En cours de publication dans *MIE2006*.
[16] Kohler F., Toussaint E.. La T2A, les pôles et la contractualisation interne. Quelles modèles en hospitalisation de court séjour?. *JFIM2005*.
[17] POMR et gestion orientée problème du dossier médical. *FULMEDDNO*. Sept 2004.

Adresse de correspondance

Professeur Darmoni Stefan
Equipe CISMeF, CHU de Rouen
Laboratoire LITIS CNRS 2645
1 rue de Germont
76031, Rouen
E-mail: stefan.darmoni@chu-rouen.fr