

Exploitation de la distance sémantique pour la création de groupements de termes en pharmacovigilance

Marie Dupuch¹, Anne Jamet^{1,2}, Marie-Christine Jaulent¹, Reinhard Fescharek³, Natalia Grabar⁴

¹Centre de Recherche des Cordeliers, Université Pierre et Marie Curie - Paris6, Paris, F-75006 ;
INSERM, U872, Paris, F-75006 France

²Centre Régional de Pharmacovigilance, HEGP, AP-HP, Paris, France

³CSL Behring GmbH, Marburg, Germany

⁴CNRS UMR 8163 STL, Université Lille 3, France

Abstract

Pharmacovigilance is the activity related to the collection, analysis and prevention of adverse drug reactions (ADRs) induced by drugs or biologics. Beside other methods, statistical methods are in use to detect new ADRs in the framework of quantitative signal detection. Groupings of terms containing similar ADRs allow to increase the signal intensity and to improve the detection of new ADRs. SMQs have become reference groupings in the field of pharmacovigilance. They are built on the MedDRA terminology and thanks to the study of scientific literature. Even if SMQs are built manually by experts, they still show some shortcomings : they tend to be over-inclusive and thus become too sensitive, although they can also miss several relevant terms. Moreover, the spectrum of available SMQs is limited. The objective of this work is to propose an automated method for a flexible creation of groupings of terms. This method is based on exploitation of the semantic distance between MedDRA terms. In a first experience, we use ADR terms alone and obtain results with a high precision (mean 74% within the interval [49 ; 91]). In a second experience, we use ADR terms and their formal definitions and this worsens the results because the semantic information within definitions may be missing. We assume that more exhaustive definitions will have a positive effect on results.

Keyword: Pharmacovigilance, signal detection, grouping of terms, SMQs, semantic distance, formal definitions

1 Introduction

La pharmacovigilance désigne l'ensemble des activités liées au recueil, à l'analyse et à la prévention des effets indésirables susceptibles d'être dus à un médicament. Non seulement pour les notifications réglementaires, mais dans pratiquement toutes les banques de données de pharmacovigilance, les effets indésirables sont codés avec les termes provenant des terminologies dédiées, WHO-ART (World Health Organization - Adverse Reaction Terminology) et MedDRA (Medical Dictionary for Drug Regulatory Activities). La recherche de signaux – relation inattendue ou encore mal définie entre un médicament et un effet indésirable – exploite ce codage et permet de détecter de nouveaux effets indésirables. Des méthodes statistiques sont typiquement utilisées pour la détection de

signaux dans des banques de données contenant un nombre suffisamment important de cas [1,2]. Mais il a été également noté que, pour augmenter l'intensité du signal, le regroupement des cas similaires d'effets indésirables suspectés est nécessaire [3]. Ce besoin devient encore plus pertinent du fait que la structure de MedDRA est très granulaire (comparée à celle de WHO-ART) : le choix des termes très précis pour le codage des effets indésirables peut avoir un effet de dilution du signal. En effet, une recherche de cas de pharmacovigilance effectuée avec un terme très spécifique peut passer à côté des cas pertinents codés avec des termes conceptuellement proches [4]. Pour contourner cette difficulté, différents niveaux hiérarchiques de MedDRA (SOC, HLT, PT) et plus récemment les SMQ (Standardised MedDRA Queries) sont utilisés. Les PT (Preferred Terms), très souvent utilisés pour la détection d'un signal, correspondent principalement à des effets indésirables très spécifiques. Tandis que les SOC (System Organ Class) et HLT (High Level Terms) correspondent à des niveaux hiérarchiques supérieurs aux PT dans MedDRA et proposent de ce fait des groupements hiérarchiques de PT relatifs à une catégorie plus générale. Cependant, l'hétérogénéité des catégories utilisées dans ces différents niveaux hiérarchiques (SOC, HLT) est une des particularités de MedDRA. Quant aux SMQ, ils regroupent tous les termes associés à une affection médicale ou un domaine d'intérêt défini. Les SMQ sont définis par des groupes d'experts grâce à l'étude manuelle de la structure de MedDRA et de la littérature scientifique [5]. Il s'agit d'un travail long et méticuleux. Quelques évaluations de ces groupements dans la détection du signal ont montré que : (1) les SMQ montrent souvent la meilleure sensibilité [6,7] mais peuvent être trop inclusifs [7] et, comme les cas retrouvés manquent de spécificité, l'évaluation des résultats par un expert médical est nécessaire et extrêmement longue. (2) Mais en même temps, le nombre d'alertes est plus élevé lorsque les PT sont utilisés [7,8]. (3) Finalement, plusieurs PT pertinents peuvent ne pas être inclus dans les SMQ [7], et (4) le répertoire des SMQ n'est pas exhaustif : plusieurs affections médicales graves n'ont pas encore été traitées par l'organisme qui fournit les SMQ.

Pour que le processus de création de groupements de termes soit plus aisé et systématique (pas d'oubli de termes), mais aussi plus optimisé (pas de groupements trop inclusifs), des méthodes automatiques peuvent être utilisées. Ces méthodes permettraient aussi aux experts d'aborder la création de nouveaux SMQ sans une connaissance *a priori* d'une terminologie. Des travaux ont été consacrés à la création automatique de groupements hiérarchiques des effets indésirables [9], mais le raisonnement hiérarchique ne respecte pas forcément le raisonnement médical dans la recherche des termes relatifs à une affection médicale. Par exemple, pour les atteintes rénales, en plus de termes comme *Acute nephritis* et *Insufficiency renal*, qui se trouvent en relation hiérarchique, il est également pertinent de prendre en considération des termes relatifs aux résultats de laboratoire (*Blood creatinine increased*, *Clearance urea decreased*) ou aux procédures médicales (*Haemodialysis*, *Peritoneal dialysis*). Nous supposons que l'exploitation de la distance sémantique (voir sec. 3) peut conduire à la création de groupements qui sont plus en accord avec le raisonnement médical. Les mesures de similarité sémantique ont été appliquées à des sous-ensembles de termes provenant des terminologies MedDRA [10] et WHO-ART [11], où les groupements obtenus montrent neuf types de relations : (1) synonymes, (2) antonymes, (3) fonctions ou anomalies physiologiques, (4) symptômes associés, (5) tests de laboratoire anormaux, (6) pathologies et leurs causes, (7) localisations anatomiques proches, (8) degré de sévérité, et (9) des groupements hétérogènes. L'évaluation de spécificité ou de sensibilité par rapport aux groupements existants (SMQ, SOC ou HLT) n'a pas été effectuée.

Dans notre travail, nous proposons de continuer d'adapter les approches de distance sémantique pour la création des groupements de termes de pharmacovigilance. L'ensemble des termes MedDRA sera utilisé. Par ailleurs, un effort spécial sera fait pour effectuer une

évaluation par rapport aux SMQ, groupements de termes établis comme une référence dans le domaine de pharmacovigilance. Dans la suite de ce travail, nous présentons d'abord le matériel exploité (sec. 2) et la méthode proposée (sec. 3). Nous présentons et discutons les résultats (sec. 4) et concluons avec quelques perspectives (sec. 5).

2 Matériel

Dans ce travail, le matériel est issu de la terminologie MedDRA [12], conçue pour coder les effets indésirables liés aux médicaments. Elle contient un spectre assez large de termes (signes et symptômes, diagnostics, indications thérapeutiques, investigations complémentaires, procédures médicales et chirurgicales, antécédents médicaux, chirurgicaux, familiaux et sociaux). Comme nous l'avons mentionné, MedDRA est organisé hiérarchiquement, et ses termes sont structurés en cinq niveaux hiérarchiques (du plus haut au plus bas) :

- 26 classes organes ou *System Organ Class* (SOC)
- 332 termes de groupes de haut niveau ou *High Level Group Term* (HLGT)
- 1 688 termes de haut niveau ou *High Level Term* (HLT)
- 18 209 termes préférés ou *Preferred Terms* (PT)
- 66 587 termes de bas niveaux ou *Low Level Terms* (LLT)

Pour le codage des effets indésirables, les autorités, l'industrie pharmaceutique et les centres de pharmacovigilance utilisent les termes MedDRA de niveau PT et LLT.

Ontologie ontoEIM. L'ontologie des effets indésirables ontoEIM [9] a été construite grâce à l'appariement des terminologies de pharmacovigilance WHO-ART et MedDRA avec la terminologie Snomed CT [13], ceci au travers de l'exploitation d'UMLS [14]. Jusqu'à 85 % de termes WHO-ART sont ainsi appariés, par contre l'appariement des termes MedDRA est beaucoup plus faible (46 %). Dans ce travail, nous exploitons la ressource construite avec les termes MedDRA. Grâce à son appariement avec Snomed CT, la représentation des termes MedDRA est enrichie de deux manières : (1) la structuration des termes est améliorée et devient parallèle à celle de Snomed CT, et (2) les termes MedDRA reçoivent des définitions formelles. Les termes peuvent ainsi être définis sur quatre axes de Snomed CT : morphologie, topographie, causalité et expression. L'exemple ci-dessous montre justement la définition sur ces quatre axes pour le concept *Arsenical keratosis* :

- morphologie : *Squamous cell neoplasm, Morphologically abnormal structure*
- topographie : *Skin structure, Structure of skin and or surface epithelium*
- causalité : *Arsenic AND OR arsenic compound*
- expression : *Abnormal keratinization*

Les termes MedDRA, leur structuration et définitions formelles dans ontoEIM sont notre matériel principal que nous utilisons pour créer des groupements des effets indésirables.

SMQ (Standardised MedDRA Queries). Les SMQ sont des groupements de termes MedDRA liés à une affection médicale donnée, comme par exemple *Acute renal failure, Hepatic disorders, Severe cutaneous reaction*. Les SMQ sont créés pour apporter une aide dans la recherche de cas pertinents en rapport avec cette affection médicale. Leur création est un processus manuel réalisé par des experts qui s'appuient d'une part sur une exploration hiérarchique (ascendante et descendante) de MedDRA et d'autre part sur la littérature scientifique. Il existe actuellement 82 SMQ, qui contiennent les termes de niveaux PT et LLT. Les SMQ existent en deux versions : *broad* et *narrow*. Dans ce travail, les SMQ en version *broad* sont utilisés comme le *gold standard* pour l'évaluation des groupements de termes.

3 Méthode

Nous proposons d'exploiter la distance sémantique pour créer des groupements de termes sémantiquement proches. La distance sémantique entre termes est le plus souvent calculée dans des structures arborescentes, comme les terminologies ou ontologies. Elle dépend du nombre d'arcs (liens) qui se trouvent entre deux termes. Par exemple, dans Snomed CT, *Abdominal abscess* se trouve à la distance de 1 de *Disorder of abdomen* et à la distance de 2 de *General finding of abdomen*. *Abdominal abscess* est donc sémantiquement plus proche de *Disorder of abdomen*. En plus du nombre d'arcs, d'autres critères peuvent intervenir. Notre méthode est organisée en trois grandes étapes : (1) calcul de la distance sémantique entre les termes et leur groupement ; (2) optimisation du niveau de termes à exploiter (PT et LLT) ; (3) évaluation des groupements par rapport aux SMQ.

3.1 Calcul de la distance sémantique et groupement de termes

La distance sémantique est calculée entre tous les termes MedDRA au sein de l'ontoEIM. Nous utilisons l'approche de Rada [15], car elle accepte la pondération des arcs. Les termes de niveaux PT et LLT sont exploités, car ils apparaissent dans les SMQ. La distance sémantique est calculée entre les termes qualifiant les effets indésirables (le plus souvent appartenant à l'axe Diagnostic *D*) et entre les éléments de leurs définitions formelles. Plus particulièrement, au sein des définitions formelles, nous exploitons les éléments provenant des axes morphologie *M* (type d'atteinte) et topographie *T* (localisation anatomique). Ces deux axes sont souvent impliqués dans les définitions des diagnostics [16] et ils sont également les plus fréquemment renseignés dans l'ontoEIM. Quant aux deux autres axes (causalité *C* et expression *E*), ils ont une apparition assez ponctuelle dans l'ontoEIM : nous ne pouvons pas nous baser sur ces derniers dans le calcul de la distance sémantique. La motivation principale pour combiner ces trois axes est donc la complétude et la précision de la méthode. Si l'axe des diagnostics *D* seul est exploité, la distance sémantique est complètement dépendante de la structure hiérarchique d'une terminologie. Afin de corriger ou d'affiner cette structuration, les définitions formelles peuvent être exploitées [17]. Pour illustrer cette approche, considérons deux termes de l'axe Diagnostic *D* *Abdominal abscess* et *Pharyngeal abscess* définis comme ceci :

- *Abdominal abscess* : $M = Abscess\ morphology$, $T = Abdominal\ cavity\ structure$
- *Pharyngeal abscess* : $M = Abscess\ morphology$, $T = Neck\ structure$

La figure 1 illustre comment sont calculées les distances (ou les plus courts chemins (pcc)) entre ces termes (axe *D*) et entre les éléments de leurs définitions formelles (axes *T* et *M*). Le poids des arcs est fixé à 1, et la valeur de chaque pcc correspond à la somme des poids de ses arcs. Pour cette paire de termes nous obtenons : $pcc_D = 4$, $pcc_T = 10$ et $pcc_M = 0$.

Pour le calcul de la distance sémantique proprement dite, les poids pcc de chaque axe sont pondérés [18]. Nous considérons ainsi que l'axe morphologie *M* est le facteur essentiel car il spécifie le type de l'atteinte morphologique (abcès, inflammation, segmentation ...) et peut fournir des indices importants pour le groupement des termes relatifs à une affection médicale. Tandis que la localisation anatomique reste secondaire et fournit, à notre avis, des indices moins importants. La distance sémantique est donc calculée de manière suivante :

$$dist_{ontoEIM}(A, B) = \frac{\sum_{i \in \{D, M, T\}} W_i * pcc(A_i, B_i)}{\sum_{j \in \{D, M, T\}} W_j}$$

où A et B sont les deux termes traités, $\{D, M, T\}$ correspondent respectivement aux axes

diagnostics, morphologie et topographie, W est le coefficient associé à chaque l'axe, et pcc est la longueur du plus court chemin sur cet axe. Les coefficients sont les suivants : $W_M=2$, $W_T=1$, $W_D=1$. C'est donc M qui reçoit le plus d'importance. La semi-matrice, construite grâce à l'application de cette formule, est exploitée par la classification hiérarchique ascendante [18] pour créer des groupements de termes. Le seuil maximum (fixé à 2) est appliqué. Le terme le plus attrayant et central d'un groupement donne le nom à ce groupement. Nous effectuons deux expériences, où nous exploitons (1) l'axe diagnostic D seul et (2) trois axes (D et deux axes des définitions formelles M et T).

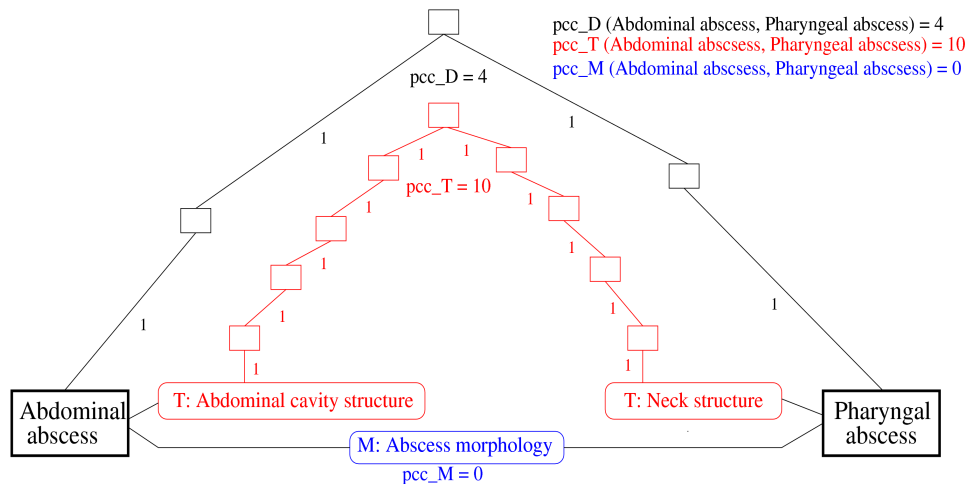


Figure 1: Les plus courts chemins (pcc) entre deux termes de niveau PT (Abdominal abscess et Pharyngeal abscess) calculés à partir de trois axes : axes des diagnostics (D) et deux axes des définitions formelles (T et M).

3.2 Optimisation du niveau de termes à exploiter (PT et LLT)

Si l'ontoEIM contient l'intégralité des termes MedDRA, seulement 46 % de ces termes ont pu être appariés avec les termes de Snomed CT. Si nous regardons de plus près l'appariement des PT et des LLT, il est respectivement de 51.6 % ($n=9\ 393$) et 35.4 % ($n=30\ 785$). Étant donné cet appariement assez partiel, nous proposons de l'optimiser pour exploiter au mieux l'ontoEIM et les définitions formelles. Nous avons pu distinguer quatre cas de figure que nous présentons ci-dessous. Dans le traitement de ces cas de figure, nous prenons en compte le fait que les termes LLT se trouvent toujours à un niveau hiérarchique inférieur par rapport aux termes PT : le pcc va donc augmenter si nous descendons d'un terme PT à un terme LLT, et diminuer si nous remontons. Même si la sémantique de la relation PT-LLT n'est pas homogène (les LLT peuvent désigner des termes synonymes, des variantes lexicales ou des termes plus spécifiques par rapport aux PT), comme dans l'ensemble de la méthode, le poids de cet arc est fixé à 1. Dans cet aspect de la méthode, nous exploitons donc la modélisation de l'ontoEIM (hiérarchie et définitions formelles) mais aussi la hiérarchie MedDRA pour la relation PT-LLT. Les quatre cas distingués sont :

- $LLT_{APPARIÉ} \in PT_{APPARIÉ}$. Les termes des deux niveaux considérés, PT et LLT, sont appariés avec les termes correspondants de Snomed CT. Le calcul de la distance sémantique peut être directement appliqué aux termes LLT et PT et à leurs définitions formelles (fig. 2(a)).
- $LLT \neg APPARIÉ \in PT_{APPARIÉ}$. Le terme de niveau PT est apparié avec un terme de Snomed CT, alors que le terme de niveau LLT n'est pas apparié. Dans ce cas, nous transposons la définition formelle du PT sur chaque LLT non apparié, et nous ajustons le pcc en ajoutant 1 à la valeur du pcc vers le PT (fig. 2(b)).
- $LLT_{APPARIÉ} \in PT \neg APPARIÉ$. Le terme de niveau LLT est apparié avec un terme de Snomed CT, alors que le terme de niveau PT n'est pas apparié. Dans ce cas, nous

transposons la définition formelle du LLT sur le terme PT. Lors du calcul du *pcc* vers le PT, nous soustrayons 1 à la valeur du *pcc* du LLT (fig. 2(c)). Comme cas particulier, un PT non défini peut avoir un seul de ses LLT défini. Dans ce cas, la définition formelle est transposée au PT (avec l'incrémentement du *pcc*) et aux autres LLT (fig. 2(d)).

– $LLT \rightarrow APPARIÉ \in PT \rightarrow APPARIÉ$. Aucun terme des deux niveaux considérés, PT ou LLT, n'est apparié avec des termes correspondants de Snomed CT. Le calcul de la distance sémantique ne peut pas être appliqué.

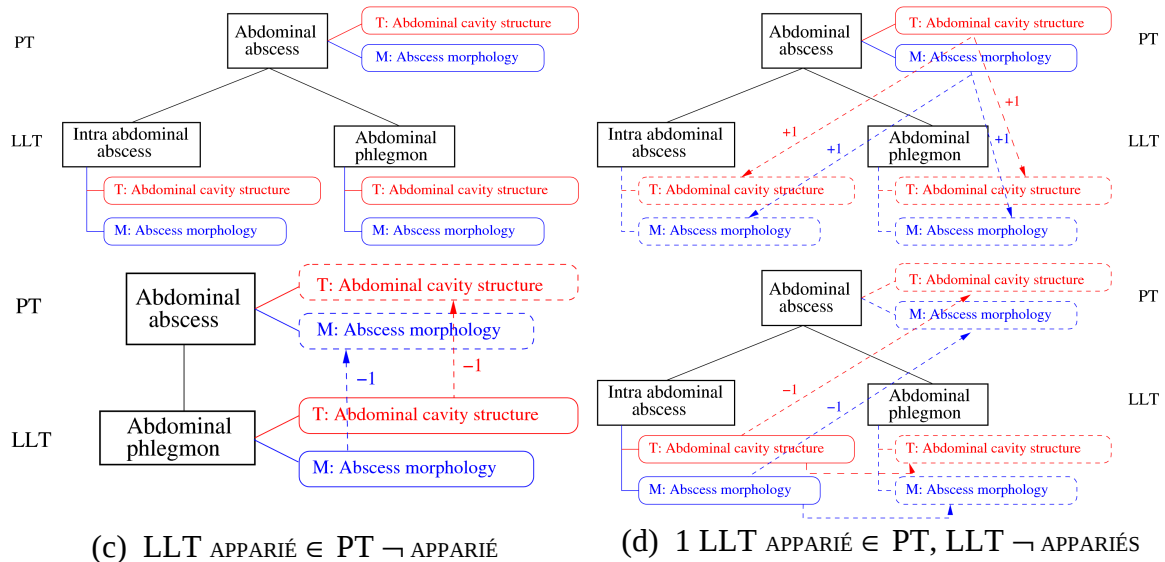


Figure 2: Optimisation de l'appariement des termes PT et LLT (en couleur noire) : transposition des définitions formelles vers les termes non appariés et ajustement des *pcc*. Les traits noirs correspondent à la hiérarchie PT-LLT de MedDRA, les traits de couleur correspondent à la structuration de Snomed CT (définitions formelles : morphologie M et topographie T).

3.3 Évaluation des groupements de termes

L'évaluation des groupements est réalisée grâce à leur comparaison avec les SMQ. Plus particulièrement, nous considérons les groupements qui correspondent à 13 effets indésirables potentiellement graves, qui se trouvent souvent à l'origine d'hospitalisations, de mise en jeu du pronostic vital ou même de décès [19]. Neuf SMQ correspondant à ces effets indésirables ont été sélectionnés par un expert (quatre autres de ces effets indésirables n'ayant pas de SMQ correspondants) : *Acute renal failure*, *Agranulocytosis*, *Anaphylactic reaction*, *Cytopenia and haematopoietic disorders affecting more than one type of blood cell*, *Gastrointestinal haemorrhages*, *Peripheral neuropathy*, *Rhabdomyolysis*, *Severe cutaneous adverse reaction*, et *Thrombocytopenia*. Ces neuf SMQ sont donc utilisés comme *gold standard* dans cette expérience. Une évaluation quantitative est effectuée avec trois mesures classiques : précision *P* (pourcentage de termes pertinents retrouvés rapporté au nombre de termes total groupés), rappel *R* (pourcentage de termes pertinents retrouvés rapporté au nombre de termes dans un SMQ) et F-mesure *F* (la moyenne harmonique de *P* et *R*). Notons que c'est aussi la F-mesure qui sert d'indice pour la mise en correspondance des SMQ avec les groupements qui leur sont les plus proches : ils ont la F-mesure la plus élevée. Une évaluation qualitative supplémentaire est réalisée avec un expert : une analyse détaillée du bruit observé dans les groupements et un jugement sur les noms des groupements.

4 Résultats et Discussion

La méthode proposée a été appliquée aux termes MedDRA au travers de la ressource ontoEIM. La distance sémantique a été calculée de deux manières : (1) avec l'axe *D*

seulement, et (2) avec trois axes (*D* et deux axes des définitions formelles *M* et *T*). Dans l'expérience avec l'axe *D*, nous obtenons 34 980 groupements. Le nombre moyen de termes par groupement est de 21,10, alors que l'intervalle varie entre 1 et 2 583 termes. Pour chaque expérience, des groupements de termes MedDRA ont été créés et évalués.

4.1 Calcul de la distance sémantique et génération des groupements avec l'axe *D*

Tableau 1 : Résultats d'évaluation des groupements par rapport aux SMQ les plus proches. Le groupement est effectué en exploitant l'axe diagnostic *D* seulement

SMQ	Groupement	Nb _t SMQ	Nb _t gr	Nb _t com	Gain (%)	P (%)	R (%)	F
Acute renal failure _c	Tubulointerstitial nephritis	267	33	19	19	57	7	12
Acute renal failure _a	Tubulointerstitial nephritis	132	33	16		48	12	19
Agranulocytosis _c	Aplasia pure red cell	308	22	17	70	77	5	10
Agranulocytosis _a	Aplasia pure red cell	151	22	10		45	6	11
Anaphylactic reaction _c	Skin oedema	707	81	40	54	49	5	10
Anaphylactic reaction _a	Skin oedema	336	81	26		32	7	12
Cytopenia	Aplasia pure red cell	120	22	19	58	86	15	26
Cytopenia	Aplasia pure red cell	54	22	12		54	22	31
Gastrointestinal haemorrhages _c	Upper gastrointestinal haemorrhage	392	44	34	22	77	8	15
Gastrointestinal haemorrhages _a	Thrombocythaemia	216	48	28		58	12	21
Peripheral neuropathy _c	Gait spastic	444	37	34	22	91	7	14
Peripheral neuropathy _a	Gait spastic	237	37	28		75	11	20
Rhabdomyolysis _c	Renal failure	319	36	29	61	80	9	16
Rhabdomyolysis _a	Renal failure	127	36	18		50	14	22
Severe cutaneous adverse reaction _c	Conjunctivitis bacterial	317	34	26	30	76	8	14
Severe cutaneous adverse reaction _a	Conjunctivitis bacterial	144	34	20		58	13	22
Thrombocytopenia _c	Thrombocythaemia	43	15	8	33	53	18	27
Thrombocytopenia _a	Thrombocythaemia	19	15	6		40	31	35

Le tableau 1 présente les résultats d'évaluation des groupements obtenus avec l'axe *D* seulement. Dans les deux premières colonnes, nous indiquons les SMQ et les groupements qui leur sont les plus proches (par leur contenu). Les trois colonnes qui suivent indiquent respectivement : le nombre de termes dans un SMQ Nb_t SMQ, le nombre de termes dans le groupement Nb_t gr et le nombre de termes communs à ces deux ensembles Nb_t com. Les trois dernières colonnes indiquent les valeurs de la précision *P*, du rappel *R* et de la F-mesure *F*. Comme nous l'avons indiqué auparavant, tous les termes MedDRA ne sont pas appariés avec les termes correspondant de Snomed CT. Nous avons décidé de prendre en compte cet aspect lors de l'évaluation, car si un terme n'est pas apparié avec Snomed CT, nous ne disposons pas d'informations nécessaires (complètes ou partielles) pour le calcul de la distance sémantique. Ainsi, pour chaque SMQ, nous considérons le SMQ complet (indexé *c*, par exemple *Acute renal failure_c*) et le même SMQ avec seulement les termes appariés avec Snomed CT (indexé *a*, par exemple *Acute renal failure_a*) et pour lesquels nous pouvons calculer les distances vers d'autres termes. Dans ce dernier cas, la taille des SMQ est ajustée et indique uniquement les termes appariés. Dans la colonne Nb_t SMQ, la différence entre les chiffres *a* et *c* indique que le pourcentage des termes appariés varie entre 39 et 55 %, ce qui laisse de côté plus de la moitié des termes dans chaque SMQ.

La taille des groupements obtenus est toujours plus petite, et leur contenu est donc plus spécifique, comparé aux SMQ. La taille reste stable indépendamment du fait que les termes pris en compte soient appariés ou non (colonne Nb_t gr), sauf pour le SMQ *Gastrointestinal haemorrhages*, où le nombre de termes diminue si l'on considère uniquement les termes appariés. Par contre, le fait de passer de la totalité des termes vers les termes appariés seulement a une influence sur le nombre de termes communs entre le groupement et le SMQ (colonne Nb_t com). Nous pouvons voir plus particulièrement qu'avec la considération des termes appariés seulement, le nombre de termes communs

avec les SMQ diminue systématiquement. Cela indique que notre méthode permet de récupérer et de traiter efficacement plusieurs termes qui ne sont pas appariés, ceci grâce à l'optimisation des appariements des termes MedDRA des niveaux PT et LLT et le calcul supplémentaire des chemins vers des termes non appariés (sec. 3.2). En considérant le rapport qui existe entre les chiffres a et c dans la colonne $Nb_t\ com$, nous obtenons un gain en termes appariés entre 20 % et 70 %. Nous considérons que ceci est un résultat très positif de la méthode.

Quant aux résultats de l'évaluation, nous pouvons remarquer que la précision est particulièrement satisfaisante. Pour les neuf paires SMQ/groupements considérées, la précision varie entre 32 % (SMQ *Anaphylactic reaction*) et 91 % (SMQ *Peripheral neuropathy*). Notons que pour sept SMQ, la précision est supérieure à 75 %. Le rappel reste faible, mais peut aller jusqu'à 31 % pour *Thrombocythaemia*. Pour certains groupements évalués ici, la F-mesure peut être considérée comme acceptable. De manière générale, la réduction des termes dans les SMQ et les groupements aux termes appariés seulement diminue la précision mais augmente le rappel et la F-mesure. Voici les moyennes et les intervalles observés :

- $P_c=73,8$ avec l'intervalle [49 ; 91] ; $P_a=51,1$ avec l'intervalle [32 ; 75]
- $R_c=9,1$ avec l'intervalle [5 ; 18] ; $R_a=14,2$ avec l'intervalle [6 ; 31]
- $F_c=16,0$ avec l'intervalle [10 ; 27] ; $F_a=21,4$ avec l'intervalle [11 ; 35]

La méthode est satisfaisante pour la précision, mais beaucoup moins pour le rappel et la F-mesure.

4.2 Calcul de la distance sémantique et génération des groupements avec trois axes

Tableau 2 : Résultats d'évaluation des groupements par rapport aux SMQ. Le groupement est effectué en exploitant trois axes : D et deux axes des définitions formelles (T et M).

SMQ	Groupement	Nb SMQ	Nb gr	Nb com	Gain (%)	P (%)	R (%)	F
Acute renal failure _c	Renal impairment	267	52	28	33	53	11	17
Acute renal failure _a	Renal impairment	132	52	21		40	16	22
Agranulocytosis _c	Aplastic anaemia	308	80	17	42	21	5	8
Agranulocytosis _a	Streptococcal sepsis	151	41	12		29	7	12
Anaphylactic reaction _c	Urticaria	707	72	32	78	44	4	8
Anaphylactic reaction _a	Urticaria	336	72	18		25	5	8
Cytopenia	Aplastic anaemia	120	80	19	137	23	15	19
Cytopenia	Aplastic anaemia	54	29	8		27	14	19
Gastrointestinal haemorrhages _c	Gastric ulcer haemorrhage	392	78	30	36	38	7	12
Gastrointestinal haemorrhages _a	Gastrointestinal haemorrhage	216	97	22		22	10	14
Peripheral neuropathy _c	Paraesthesia	444	54	47	67	87	10	18
Peripheral neuropathy _a	Gait spastic	237	45	28		62	11	19
Rhabdomyolysis _c	Muscular weakness	319	81	24	140	30	7	12
Rhabdomyolysis _a	Muscular weakness	127	52	10		19	7	11
Severe cutaneous adverse reaction _c	Skin exfoliation	317	102	27	42	26	8	12
Severe cutaneous adverse reaction _a	Skin exfoliation	144	102	19		18	13	15
Thrombocytopenia _c	Platelet count abnormal	43	21	5	0	23	11	15
Thrombocytopenia _a	Platelet count abnormal	19	21	5		23	26	25

Le tableau 2 présente les résultats globaux d'évaluation des groupements obtenus avec trois axes D , M et T : il prend en compte l'information sur les termes et sur leurs définitions formelles. Sa structure et ses informations sont similaires à celles du tableau 1. Comme dans les résultats précédents, la taille des groupements est plus petite que la taille des SMQ. Par contre, en fonction des types de termes (appariés ou non), la taille d'un même groupement (colonne $Nb_t\ gr$) peut varier. Ce fait transparait aussi au travers du nombre de termes communs (colonne $Nb_t\ com$) entre le groupement et le SMQ. Il devient ici plus saillant que dans les résultats du tableau 1. Nous pouvons ainsi voir qu'avec la

considération des termes appariés uniquement, le nombre de termes communs avec les SMQ diminue systématiquement, sauf pour le SMQ *Thrombocytopenia*. Comme auparavant, le gain en termes appariés (sauf pour le SMQ *Thrombocytopenia*) est important. Il est toujours supérieur à 30 %. Il atteint jusqu'à 140 % pour le SMQ *Rhabdomyolysis*, où nous passons de 10 termes appariés communs à 24 termes communs, dont 14 ont des informations sémantiques inférées.

En ce qui concerne les mesures d'évaluation, nous pouvons voir que si globalement le rapport entre ces mesures reste le même (la précision est la plus élevée et le rappel le plus faible), les résultats sont inférieurs à ceux présentés dans le tableau 1. Cela concerne les trois mesures considérées : précision, rappel et F-mesure. Dans une certaine mesure, cette observation est contraire à nos attentes. Effectivement, nous pensions que des informations sémantiques plus riches (c'est-à-dire la prise en compte des définitions formelles) allaient améliorer les résultats. Comme nous allons le montrer dans la suite de ce travail (sec. 4.3), les définitions formelles sur lesquelles nous nous basons sont incomplètes, ce qui biaise les distances sémantiques et donc les groupements. Voici les moyennes et les intervalles que nous obtenons lors de cette deuxième expérience :

- $P_c=38,3$ avec l'intervalle [21 ; 87] ; $P_a=29,4$ avec l'intervalle [18 ; 62]
- $R_c=8,7$ avec l'intervalle [4 ; 15] ; $R_a=12,1$ avec l'intervalle [5 ; 26]
- $F_c=13,4$ avec l'intervalle [8 ; 19] ; $F_a=16,1$ avec l'intervalle [8 ; 25]

Comme auparavant, la réduction des SMQ complets vers des SMQ avec les termes appariés seulement diminue la précision mais augmente le rappel et la F-mesure.

4.3Analyse détaillée

Nous analysons maintenant plus en détail quelques groupements. Nous considérons le SMQ *Peripheral neuropathy* du tableau 1 (axe *D* seulement), et deux SMQ du tableau 2 (trois axes) : *Acute renal failure* et *Anaphylactic reaction*. Nous nous concentrons surtout sur le bruit généré par la méthode.

Le SMQ *Peripheral neuropathy* (444 termes) est associé au groupement *Gait spastic* (37 termes). Ils ont 34 termes communs. Les trois faux positifs de ce groupement sont : *Difficulty in walking involving joint site unspecified*, *Wide based gait* et *Gait spastic*. Les deux premiers sont des LLT du PT *Gait disturbance*, qui lui, est inclus dans le SMQ. Ce PT *Gait disturbance* possède 35 LLT qui, pour un grand nombre d'entre eux, appartiennent à cet SMQ. Concernant le troisième terme (un PT), il correspond au nom du groupement et se trouve donc sémantiquement central. Selon la structure de ontoEIM, ces trois termes sont potentiellement pertinents pour le SMQ *Peripheral neuropathy*.

Le SMQ *Acute renal failure* (267 termes) est associé au groupement *Renal impairment*, dont le terme central est défini sur l'axe topographie uniquement ($T = \textit{Kidney structure}$). L'axe morphologie de ce terme n'est pas renseigné et donc ne peut pas être pris en compte dans le calcul de la distance sémantique. Le groupement *Renal impairment* contient 52 termes, dont 28 sont communs avec le SMQ, ce qui donne les chiffres d'évaluation suivantes : $P=0,53$, $R=0,11$, $F=0,17$. Nous allons analyser en détail le bruit généré dans ce groupement. Parmi les 24 termes qui sont réunis dans le groupement mais qui ne se trouvent pas dans le SMQ, nous avons :

- 14 termes avec des définitions formelles partielles : seulement l'axe *T* est défini et l'axe *M* est absent. C'est le cas du terme central *Renal impairment*. Il concerne aussi d'autres termes comme *Renal cancer*, *Renal aplasia*, *Acute diffuse glomerulonephritis*, *Renal agenesis congenital*. Le trait commun entre ces termes, qui est la topographie, permet leur regroupement, alors que du point de vue de la morphologie, il s'agit d'atteintes très différentes (cancer, aplasie, inflammation ...).

- 5 termes dont les définitions formelles ne sont pas renseignées et pour lesquels seul le diagnostic est présent. Cela concerne les termes comme *Autonomic failure syndrome* et *Type V hyperlipidaemia*.
- 1 cas où plusieurs chemins sont possibles et où le plus court chemin est inférieur au seuil fixé. Il s'agit de *Post procedural complication*.
- Finalement, les 4 termes qui restent (*Renal agenesis and dysgenesis*, *Single functional kidney*, *Kidney absent* et *Renal agenesis*) sont tous sémantiquement très proches du terme principal et, selon l'avis de l'expert, sont pertinents pour le SMQ *Acute renal failure*.

Et un dernier exemple : SMQ *Anaphylactic reaction* (707 termes) associé au groupement *Urticaria* (72 termes). Le terme central du groupement *Urticaria* est défini sur deux axes *M* (*M* = *Urticaria*) et *T* (*T* = *Skin structure*). Ce groupement a seulement 18 termes communs avec le SMQ. Si le groupement contient autant de bruit ($n=54$), ceci est dû essentiellement à deux facteurs : (1) 16 termes du groupement ne sont définis que sur l'axe *T* (i.e., *Dermatitis radiation*, *Porphyria non acute*), ce qui introduit des rapprochements biaisés et faux ; (2) un des PT *Urticaria* du SMQ possède 34 LLT, dont 18 LLT ne sont pas inclus dans le SMQ. Par contre, notre méthode les regroupe et ce regroupement pourrait être considéré comme correct.

L'analyse du bruit montre que notre méthode est très dépendante de l'exhaustivité des définitions formelles et, lorsque ces définitions ne sont pas complètes, la distance sémantique et les groupements deviennent biaisés. En effet, parmi tous les faux négatifs analysés, aucun n'est défini sur les deux axes. Cette lacune dans les définitions formelles conduit au fait que les résultats présentés dans le tableau 1 sont meilleurs que ceux du tableau 2, ce qui est contraire à nos attentes. Si ces expériences ont montré la faisabilité de l'approche et ont permis d'obtenir des résultats avec une bonne précision. Nous pensons que les définitions formelles présentent un potentiel important et permettront, lorsqu'elles seront plus exhaustives, d'améliorer la qualité des groupements.

Une évaluation supplémentaire concerne la dénomination des groupements : les noms des groupements sont majoritairement en rapport avec les noms des SMQ correspondants (i.e., {*Acute renal failure*, *Renal impairment*}, {*Gastrointestinal haemorrhage*, *Gastric ulcer haemorrhage*}), de même qu'avec les termes au sein des groupements. Mais dans quelques exemples ({*Peripheral neuropathy*, *Gait spastic*}, {*Gastrointestinal haemorrhage*, *Thrombocythaemia*}) les noms des SMQ et des groupements mis en correspondance n'ont pas beaucoup de traits sémantiques communs. Cet aspect doit être évalué plus en détail et la méthode de dénomination des groupements doit certainement être améliorée. D'autant plus que la dénomination des groupements aura une importance primordiale si ces groupements sont utilisés pour la création de nouveaux SMQ : le nom donné aux groupements devra alors être très précis pour aider les experts de les exploiter au mieux.

5 Conclusion et Perspectives

Nous avons proposé une méthode, qui exploite la distance sémantique, pour la création de groupements d'effets indésirables. Ce travail a une implication potentielle dans la détection de signaux de pharmacovigilance dans des bases de données qui augmentent sans cesse : les méthodes de ce type peuvent en effet aider à rapprocher des cas encodés avec des termes différents mais qui sont proches sémantiquement. Par ailleurs, ces méthodes peuvent aussi accélérer et faciliter la création de nouveaux SMQ ou groupements de termes. Dans notre travail, deux expériences ont été réalisées : (1) exploitation de l'axe diagnostique *D*, et (2) exploitation de trois axes (*D* et deux axes des définitions formelles : *M* morphologie et *T* topographie). Les groupements ont été évalués par rapport aux SMQ

version *broad* avec lesquels ces groupements ont la F-mesure la plus élevée. Ces expériences ont montré qu'avec l'utilisation de l'axe *D* seul les performances sont plus élevées qu'avec l'utilisation des trois axes : la précision montre des valeurs moyennes de 73,8 % et 38,3 % respectivement. Le rappel reste faible dans les deux expériences. Cette baisse de la précision vient du fait que les termes, même s'ils sont appariés avec Snomed CT, ne sont pas systématiquement ni exhaustivement définis. Par ailleurs, seulement 46 % des termes MedDRA, qui sont appariés avec Snomed CT, peuvent être traités par cette méthode. Nous avons réussi à contourner en partie cette limitation et obtenons un gain substantiel en termes appariés (entre 20 % et 140 % selon les SMQ). Si, dans ce travail, nous nous sommes concentrés sur l'analyse de la précision et du bruit, l'analyse des silences reste à faire et l'amélioration du rappel demandera encore d'autres stratégies, comme par exemple la fusion des groupements proches. D'autres perspectives concernent l'amélioration des dénominations de groupements et des expériences avec les SMQ en version *narrow*. Des expériences complémentaires peuvent conduire à l'ajustement des seuils et des variables (poids des arcs, coefficients des axes). Nous pensons aussi que les méthodes du traitement automatique de langue peuvent affiner et enrichir les groupements.

Remerciements. Ce travail a été partiellement financé par le Septième Programme Cadre de la Communauté Européenne (FP7/2007-2013) pour le programme *Innovative Medicine Initiative* (IMI), Grant Agreement n° 1150004. Les travaux de recherche qui ont produit ces résultats ont été conduits au sein du consortium PROTECT (*Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium*, www.imi-protect.eu), qui est un partenariat public-privé coordonné par la European Medicines Agency. Les auteurs remercient d'autres participants à cette tâche (C. Bousquet, O. Caster, G. Declerck, R. Hill, A. Kluczka, X. Kurz, M. Lerch, N. Noren, V. Pinkston, E. Sadou, J. Souvignet et T. Vardar), mais seuls les auteurs sont responsables des opinions exprimées ici.

Références

- [1]Bate A, Lindquist M, Edwards I, Olsson S, Orre R, Lansner A, De Freitas R. (1998). A bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol*, 54(4), 315–21.
- [2]Meyboom R, Lindquist M, Egberts A, Edwards I. (2002). Signal selection and follow-up in pharmacovigilance. *Drug Saf*, 25(6), 459–65.
- [3]Hauben M, Bate A. (2009). Decision support methods for the detection of adverse events in post-marketing data. *Drug Discov Today*, 14(7-8), 343–57.
- [4]Fescharek R, Kübler J, Elsasser U, Frank M, Güthlein P. (2004). Medical dictionary for regulatory activities (MedDRA) : Data retrieval and presentation. *Int J Pharm Med*, 18(5), 259–269.
- [5]CIOMS (August 2004). Development and Rational Use of Standardised MedDRA Queries (SMQs) : Retrieving Adverse Drug Reactions with MedDRA. Rapport interne, CIOMS.
- [6]Mozzicato P. (2007). Standardised MedDRA queries : their role in signal detection. *Drug Saf*, 30(7), 617–9.
- [7]Pearson R, Hauben M, Goldsmith D, Gould A, Madigan D, O'Hara D, Reisinger S, Hochberg A. (2009). Influence of the MedDRA hierarchy on pharmacovigilance data mining results. *Int J Med Inform*, 78(12), 97–103.

- [8]Yuen N, Fram D, Vanderwall D, Almenoff J. Do standardized MedDRA queries add value to safety data mining? In ICPE 2008.
- [9]Alecu I, Bousquet C, Jaulent MC. (2008). A case report : using Snomed CT for grouping adverse drug reactions terms. BMC Med Inform Decis Mak, 8(S1), 4.
- [10]Bousquet C, Henegar C, Louet A, Degoulet P, Jaulent MC. (2005). Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach. Int J Med Inform, 74(7-8), 563–71.
- [11]Iavindrasana J, Bousquet C, Degoulet P, Jaulent MC. (2006). Clustering WHO-ART terms using semantic distance and machine algorithms. In AMIA Annu Symp Proc, p. 369–73.
- [12]Brown E, Wood L, Wood S. (1999). The medical dictionary for regulatory activities (MedDRA). Drug Saf., 20(2), 109–17.
- [13]Stearns M, Price C, Spackman K, Wang A. (2001). Snomed clinical terms : overview of the development process and project status. In AMIA, p. 662–666.
- [14]NLM (2008). UMLS Knowledge Sources Manual. National Library of Medicine, Bethesda, Maryland. www.nlm.nih.gov/research/umls/. (2010)
- [15]Rada R, Mili H, Bicknell E, Blettner M. (1989). Development and application of a metric on semantic nets. IEEE Transactions on systems, man and cybernetics, 19(1), 17–30.
- [16]Spackman K, Campbell K. (1998). Compositional concept representation using Snomed : Towards further convergence of clinical terminologies. In Journal of American Medical Informatics Association (JAMIA), p. 740–744.
- [17]Petiot D, Burgun A, Le Beux P. (1996). Modelisation of a criterion of proximity : Application to medical thesauri. In Medical Informatics Europe, p. 149–52.
- [18]Lebart L, Salem A. (1994). Statistique textuelle. Dunod, Paris
- [19]Trifirò G, Pariente A, Coloma P, Kors J, Polimeni G, Miremont-Salamé G, Catania M, Salvo F, David A, Moore N, Caputi A, Sturkenboom M, Molokhia M, Hippisley-Cox J, Acedo C, Van Der Lei J, Fourier-Reglat A. (2009). EU-ADR group. Data mining on electronic health record databases for signal detection in pharmacovigilance : Which events to monitor? Pharmacoepidemiol Drug Saf, 18(12), 1176–84.

Adresse de correspondance

Marie Dupuch
SPIM, UMRS Inserm 872 éq. 20
Centre de Recherche des Cordeliers
15 rue de l'Ecole de Médecine
75006 Paris
marie.dupuch@crc.jussieu.fr