Un jeu de données minimum pour faciliter l'interopérabilité des bases de données pour les maladies rares

Rémy Choquet[1,2,3], Claude Messiaen[1,2,4], Adrien Priouzeau[1], Albane de Carrara[1,2], et Paul Landais[1,2,4]

[1] Groupe de travail BNDMR/ISy-Rare
[2] Assistance Publique des Hôpitaux de Paris, 75000 Paris
{remy.choquet,claude.messiaen,paul.landais,
albane.decarrara}@nck.aphp.fr
adrien.priouzeau@trs.aphp.fr
[3] Laboratoire Ingénierie des Connaissances en Santé,
UMRS872 EQ20, 75005 Paris
[4] Université Paris Descartes EA 4472, 75005 Paris

Résumé : La définition de vocabulaires ou de modèles de données communs apporte une aide importante pour garantir l'interopérabilité des systèmes d'information en santé. Dans le cadre de plans nationaux pour la mise en œuvre d'un système d'information commun à toutes les maladies rares (environ 7000), nous proposons une méthode permettant la définition d'un jeu de données minimum (minimum dataset) pour les maladies rares. La définition d'un jeu d'éléments de données commun minimum permet de s'assurer qu'un tronc commun d'élément de données partageables dans le domaine des maladies rares à l'échelon national. Nous proposons une méthodologie spécifique et nous publions les premiers résultats de notre étude que nous discutons.

Mots-clés : Minimum Dataset, Interopérabilité sémantique, Maladies rares, Jeu de données minimum

1 Introduction

Le plan national maladies rares 2 (2011-2014) promeut la mise en œuvre d'une base de données nationale maladies rares (BNDMR) pour rendre compte de la demande et de l'offre de soins dans le domaine des maladies rares. Le deuxième enjeu de la BNDMR est de permettre le recrutement de patients pour la mise en place de protocoles de recherche (cohortes). Le recueil de données s'effectue dans des centres maladies rares (CMR) disséminés sur le territoire Français. Le mode de recueil est

libre. Chaque CMR a, aujourd'hui, mis en œuvre les modalités de recueil selon ses propres moyens (dossier patient généraliste, dossier patient maladies rares, registre, cohorte, dossier papier). La nature de l'information recueillie est aussi très hétérogène puisque les maladies rares sont généralement représentées dans toutes les spécialités médicales, et deviennent même, des spécialités à part entière. On dénombre entre 6000 et 7 000 maladies rares et on estime la file active de patients entre 250 000 et 500 000 sur le territoire français. L'objectif de notre travail est la définition d'un minimum dataset (MDS) commun à toutes les maladies rares (MR). Nous proposons une méthode de mise en œuvre s'appuyant sur les littératures grises et scientifiques, et sur un comité d'experts. Nous présentons nos résultats préliminaires et nous les discutons.

2 Etat de l'art

Le concept de minimum dataset s'est développé pour mettre en œuvre des études à grande échelle dans le domaine de la santé (Sheila 2008, Pheby 1994 et Webster 1998) avec des données comparables. Des minimums datasets sont généralement proposés pour des études spécifiques, et parfois dans des domaines entiers tels que l'oncologie pour la mise en œuvre de registres[1]. Dans le domaine des maladies rares, une revue de la littérature préliminaire met en avant un minimum « common data elements » pour les maladies rares par l'office Américain des maladies rares[2]. En Europe, une initiative équivalente est étudiée dans le projet Epirare[3] mais pour les registres. En France, le programme CEMARA a développé un minimum dataset pour 55 CMR (Landais 2010).

Peu de méthodes sont proposées dans la littérature pour mettre en œuvre un mimum dataset, nous retiendrons particulièrement les travaux de (Svensson-Ranallo et al. 2011) qui offrent une méthodologie globale de mise en œuvre d'un minimum dataset que nous adapterons à notre étude.

3 Méthodologie et matériel

Nous présentons des méthodologies spécifiques associées à la constitution du MDS MR grâce à la littérature scientifique et grise. Nous décrivons en premier lieu la méthodologie générale de notre travail :

[1] http://www.ukacr.org/content/current-cancer-registry-minimum-dataset

[2] http://www.grdr.info/index.php/common-data-elements

[3] http://www.epirare.eu/

1. Définition d'un groupe de travail MDS MR et d'un groupe d'experts
2. Etude des besoins auprès des centres maladies rares
3. Etude de la littérature scientifique et grise
4. Proposition d'une première version du MDS MR pour avis et priorisation auprès des groupes d'experts
5. Revue sémantique du MDS MR et standardisation des éléments de données
6. Publication du MDS MR

Méthode de revue systématique de la littérature scientifique

L'objectif étant la revue systématique de la littérature scientifique traitant de la création d'un minimum dataset, nous avons commencé par élaborer un corpus de mots-clés pour restreindre notre recherche selon nos besoins. Les mots-clés retenus par notre groupe d'experts sont les suivants : *Dataset, Data set, Minimum dataset, Minimum data set, Data catalog, Data catalogue, Data model, Model of data, Models of data, Common data elements*.

Notre première revue a été effectuée en associant chaque mot-clé du corpus au terme « rare disease ». Les résultats de cette recherche n'ont pas été concluants. Nous avons donc développé une méthode de recherche basée sur l'ontologie d'Orphanet afin de gagner en exhaustivité. Nous avons développé un outil permettant d'utiliser un corpus de mots-clés que nous associons à une ontologie afin d'effectuer une revue systématique de la littérature dans PubMed. Ici, les 7000 maladies rares (ou groupes de maladies rares) sont associées à chaque mot-clé du corpus, par exemple : « Cystic fibrosis » et « common data elements ». L'algorithme de recherche défini est le suivant :

```
Extraction des maladies de l'ontologie d'Orphanet
    Pour chaque maladie (ou groupe de maladies),
          Pour chaque mot-clé,
                -association de la maladie au mot-clé et
                création de la requête PubMed
                -appel au WebService : transmission de la
                requête à PubMed
                -réception des résultats, un ensemble de
                PMID (PubMed IDentifier)
                -appel au WebService : transmission des
                PMID résultats pour obtenir les titres et
                résumés des articles
                -réception des résultats, titres et
                résumés des articles
                -ecriture des résultats
```

Les résultats sont ensuite revus par un expert du domaine.

Méthode de revue systématique de la littérature grise

Un questionnaire national a été mis en œuvre et proposé aux 131 centres de références maladies rares en France. Ce questionnaire nous a permis d'identifier les applications existantes dans les CMR permettant de colliger des données maladies rares lors des visites des malades. Nous avons identifié 3 applications majeures existantes en France (CEMARA, eRespiRare et ESID). Nous avons enfin cherché à identifier des MDS MR dans d'autres pays au travers des ministères habituellement en charge de ce type de projet. Nous avons colligé ces informations afin de pouvoir les aligner et proposer un premier MDS MR.

Matériel pour la constitution du MDS MR

Nous avions à notre disposition deux datasets maladies rares de deux recueils français : CEMARA[4], actuellement adopté par 55 CMR et eRespiRare, un système d'information de type dossier patient dédié aux maladies respiratoires. Nous avons également obtenu le minimum dataset de *The Office of Rare Diseases Research* (ORDR)[5], département américain dédié aux maladies rares, et de The European Society for Immunodeficiencies (ESID)[6] (Guzman 2007). Par ailleurs, un projet Européen de mise en œuvre d'un minimum dataset pour les registres en Europe est actuellement en cours (EpiRare[7]), mais aucun dataset n'est à ce jour publié.

4 Résultats

Nous présentons ici les résultats issus de la revue de la littérature scientifique et de la littérature grise pour la constitution d'une première version du MDS MR.

L'outil de revue de la littérature scientifique basé sur une ontologie et sur un corpus de mots-clés développé a envoyé, pour notre cas d'étude, à PubMed, environ 100 000 requêtes. En effet, nous avons environ 7 000 maladies dans l'ontologie d'Orphanet et une dizaine de mots-clés, ce qui correspond à 70 000 requêtes associant maladies et mots-clés. Pour chaque requête retournant des résultats, nous envoyons une nouvelle requête à PubMed afin de récupérer le titre et l'extrait de chaque article.

[4] CEMARA : http://cemara.org/

[5] ORDR : http://rarediseases.info.nih.gov/

[6] ESID : http://www.esid.org/

[7] EPIRARE : http://www.epirare.eu/

En supposant que chaque requête renvoie un résultat, nous envoyons donc 70 000 requêtes de plus, pour un total de 140 000 interrogations de PubMed au maximum.

En parallèle de la revue experte nous avons relancé une recherche, uniquement sur les titres des articles issus de la revue systématique (et non plus sur les titres et extraits), en n'utilisant que les mots-clés. Les résultats obtenus ont eux-aussi été soumis à un expert.

Nous avons obtenu 2 126 résultats issus de la recherche automatisée. Après revue par un expert, seul 8 résultats ont été retenus. Parmi les articles refusés, certains étaient trop spécifiques, d'autres employaient les mots-clés dans un tout autre contexte que le nôtre.

La recherche sur les titres des articles a quant à elle remonté 31 articles sur les 2 126 résultats de départ. Seuls 3 ont finalement été retenus après revue experte. Ces 3 articles font partie des 8 retenus après revue experte des résultats de la recherche sur les titres et extraits. Nous ne retrouvons donc que 37.5% des articles pertinents du corpus.

Finalement, cette comparaison nous permet d'affirmer que l'approche par spécialisation sur les titres est trop restrictive, bien que moins coûteuse pour l'expert.

Pour la constitution du premier MDS MR qui sera soumis aux comités d'experts, nous avons alignés les datasets mis à notre disposition. Nous avons d'abord aligné nos deux sources françaises avec la source américaine et la source européenne. Nous avons conservé les items communs aux différents MDS mais nous avons également conservé ceux qui étaient congruents avec les objectifs de notre démarche. A l'issu de ce double alignement, nous avons pu mettre en évidence 13 groupes de données : *identification du patient* (21 items), *communication avec le patient* (24 items), *information sur la structure de soins* (19 items), *informations médicales sur les patients* (19 items), *données anténatales et néonatales* (10 items), *diagnostic du patient* (11 items), *traitement* (4 items), *activité de soins* (8 items), *participation à la recherche* (4 items), *consentement* (4 items) , *information matériel biologique* (6 items) et aspects *sociodémographiques* (7 items).

Pour valider notre approche, nous avons aligné les datasets suggérés dans la littérature scientifique au MDS MR version 1. Nous avons tout d'abord écarté les articles jugés non pertinents par notre expert. Nous avons ensuite aligné le corpus d'articles pertinents (5 articles) au MDS MR.

5 Discussion

A l'issue de la revue systématique, l'ensemble des articles que nous avons obtenus peut être divisé en deux parties : les Vrai-Positifs (VP),

c'est-à-dire les articles correspondant réellement à notre recherche, ceux qui sont retenus après la revue experte, et les Faux-Positifs (FP), c'est-à-dire les articles remontés mais ne correspondant pas, rejetés après revue experte. La version actuelle de l'outil revue systématique ne permet pas de mettre de côté les Faux-Positifs (FP), et ainsi réduire le corpus d'articles à vérifier par les experts. D'autre part, se pose aussi la question des Faux-Négatifs (FN), articles non-renvoyés par la revue systématique mais qui correspondent à nos besoins.

Notre sélection issue de la littérature grise n'est pas exhaustive et les objectifs de ces MDS sont différents (registre, cohorte, dossier de spécialité). Cependant, nos choix se justifient par la spécificité du domaine des maladies rares et le peu de solutions disponibles.

Notre méthode générale ne sera pas validée ici car la mise en œuvre de celle-ci est toujours en cours. Nous pouvons cependant remarquer que celle-ci s'appuie sur un ensemble de ressources et d'acteurs, et qu'elle est généralisable pour la constitution de datasets plus spécifiques qui seront développés par ailleurs au sein de notre banque nationale maladies rares dans le cas de cohortes, notamment.

Au final, nous avons pu constituer une première version du minimum dataset maladies rares dans le contexte défini par le plan national maladies rares 2. Nous avons proposé ce dataset pour revue experte auprès du ministère de la santé et des centres de référence. Nous estimons que ce minimum dataset facilitera l'interopérabilité de notre système d'information maladies rares[8].

Références

BIRD S. & FARRAR J. (2008). Minimum dataset needed for confirmed human H5N1 cases. Lancet. vol. 372 (9640) pp. 696-7.

GUZMAN ET AL. (2007). The ESID Online Database network. Bioinformatics. vol. 23 (5) pp. 654-5.

LANDAIS ET AL. (2010). CEMARA an information system for rare diseases. Stud Health Technol Inform. vol. 160 (Pt 1) pp. 481-5.

PHEBY F. & ETHERINGTON J. (1994). Improving the comparability of cancer registry treatment data and proposals for a new national minimum dataset. J Public Health Med. vol. 16 (3) pp. 331-40.

SVENSSON-RANALLO ET AL. (2011). A framework and standardized methodology for developing minimum clinical datasets. AMIA Summits Transl Sci Proc. vol. 2011 pp. 54-8.

WEBSTER D. (1998). A minimum dataset for newborn screening. J Med Screen. vol. 5 (2) pp. 109.

[8] http://www.isyrare.fr