Application de l'analyse statistique des données textuelles à une revue bibliographique de la littérature médicale

Jacques Rouillier[1], Jean-Yves Bansard[2], Michel Kerbaol[2]

[1] Département de Médecine Générale – Université de Rennes 1 – Rennes – France
[2] INSERM-Département de Santé Publique – Université de Rennes 1 – Rennes – France

Abstract

With the aim of reviewing the current overhang of the medical research concerning the evaluation of physical activity in the elderly, the review of the literature is proposed by using a new statistical tool : textual data analysis. After a critical presentation of the on-line accessible data banks, the application of this methodology is experimented to identify, within a documentary corpus of 6 107 abstracts extracted from data banks Medline® and SCI®, the relevant publications. Statistical analysis allows extraction of a corpus of 143 publications. The comparison with a classic search by logical equation in the same data bases shows the sharply superior performances of our methodology : more than half of the relevant publications are ignored by the classic search by logical equation. The methodology using the data analysis allows to free the bibliographical search from the problems of "noise" and "silence", in order to gather in a reliable and exhaustive way all the bibliographical references on the transverse subject of the human physical activity evaluation. One can legally suppose in the fall of this experiment that these performances should be exploited during any bibliographical search.

Keywords: textual data analysis, correspondence analysis, bibliography, medicine

Résumé

Dans le but de faire le point sur l'avancée actuelle de la recherche médicale concernant l'évaluation de l'activité physique chez les sujets âgés, une revue de la littérature est proposée en utilisant un nouvel outil statistique : l'analyse de données textuelles. L'application de cette méthodologie est expérimentée pour identifier, au sein d'un corpus documentaire de 6 107 résumés extrait des banques de données Medline® et SCI®, les publications pertinentes. L'analyse factorielle de correspondance permet d'extraire un corpus de 143 publications. La comparaison avec une recherche classique par équation logique dans les mêmes bases de données montre les performances nettement supérieures de notre méthodologie : plus de la moitié des publications pertinentes sont ignorées par la recherche classique par équation logique. La méthodologie utilisant l'analyse de données permet d'affranchir la recherche bibliographique des problèmes de « bruit » et de « silence », pour réunir de manière fiable et exhaustive l'ensemble des références bibliographiques sur le sujet transversal de l'évaluation de l'activité physique humaine. On peut légitimement supposer au décours de cette expérimentation que ces performances devraient être exploitées au cours de toute recherche bibliographique.

Mots-clés : analyse de données textuelles, analyse de correspondance, bibliographie, médecine.

1. Introduction

De nombreux auteurs contribuent à l'avancée des connaissances médicales, ce qui a pour conséquence la production d'une abondante littérature scientifique, publiée dans les revues spécialisées et référencée dans plusieurs bases de données. Le défi initial lancé au chercheur est actuellement de retrouver l'information pertinente dans cette profusion, afin d'établir les bases de son travail. On propose ici une réponse au questionnement obligatoire et préliminaire à toute recherche : faire le point de manière satisfaisante sur l'avancée des connaissances dans le domaine considéré.

L'abondance de la littérature médicale est devenue telle qu'en extraire les publications concernant un domaine précis devient une gageure. La multiplication des bases de données bibliographiques constitue davantage un obstacle pour le chercheur qu'elle n'assure l'exhaustivité dans la compilation des publications. La recherche par équation logique dans les bases de données disponibles sur la Toile ne parvient le plus souvent qu'à une extraction biaisée par les préconçus du chercheur, la nature des bases de données, les choix de référencement de leurs responsables, l'indexation humaine des titres et des résumés, et bien entendu la politique éditoriale des publications indexées. Le résultat d'une recherche classique est donc le plus souvent partial, et surtout partiel. L'analyse de données textuelles semble pouvoir fournir une méthode satisfaisante en terme de fiabilité et d'exhaustivité, sans pouvoir compenser toutefois les défauts intrinsèques à la conception actuelle des bases de données.

2. La recherche bibliographique : procédure

Le thème choisi ici est l'activité physique chez le sujet âgé, de nombreux travaux ayant souligné les bénéfices de cette activité en termes de santé. On note qu'il s'agit d'un domaine précis mais transversal, intéressant plusieurs disciplines médicales, obstacle courant mais qui complique la recherche bibliographique. L'objectif est d'inventorier toutes les méthodes d'évaluation de l'activité physique humaine, et d'en étudier la validité pour construire un outil utilisable en pratique courante de médecine générale.

2.1. Recueil d'une base documentaire de travail

La méthode choisie consiste à recueillir un corpus documentaire large : il convient davantage à ce stade d'éviter les « silences » que les « bruits », et l'équation est destinée à recueillir à peu près sûrement toutes les publications pertinentes. Cette équation logique s'écrit :

(physical activity OR sport* OR exercise OR fitness OR training OR leisure)
AND (elderly OR older OR aging OR retired)

La recherche avec cette équation logique au sein des deux bases de données que sont MEDLINE®[1] et SCI®[2] conduit à un corpus de 6 270 fiches documentaires. Après un « raffinage » destiné à éliminer les fiches « en double » et celles ne comportant pas de résumé (donc inaccessibles à l'analyse de données textuelles), et une recherche complémentaire de documents comportant des notions intéressantes (nouvelle interrogation de la base MEDLINE® avec les mots « questionnaire » ou « évaluation »), ou encore des « related articles » associés à des fiches considérées comme très pertinentes, on aboutit à une base de travail exploitable de 6 107 documents, sur laquelle portera l'analyse de données textuelles.

2.2. Analyse des occurrences de mots

Un dictionnaire est constitué grâce au logiciel BI[3], qui référence l'intégralité des mots figurant dans les titres et les résumés de tous les documents. Ce dictionnaire est constitué de 25 017 mots différents.

Après élimination des mots « parasites », le calcul de la fréquence de chacun des mots du dictionnaire est établi successivement pour l'ensemble des 6 107 documents de la base de travail, et pour chacun de ces documents. Un tri par ordre décroissant de fréquences permet de

[1] MEDLINE® : Base de données bibliographiques biomédicales de la National Library of Medicine (NLM).
[2] SCI® : Science Citation Index, index thématique de l'Institute for Scientific Information (ISI).
[3] Logiciel conçu par Michel KERBAOL et Joël JOSSE, Log INSERM © 1979 – 1987 – 1993.

constituer un fichier des 1000 premiers mots pour établir le tableau croisant les 6107 fiches bibliographiques avec la fréquence d'occurrence de ces 1000 mots.

2.3. Analyse factorielle de correspondance

Le tableau est ensuite analysé à l'aide du logiciel ADDAD[4], et les résultats exploités avec QNOMIS-II[5]. L'analyse factorielle de correspondance (AFC) conduit à l'identification de 30 facteurs avec pour chacun d'eux une valeur positive ou négative. Chacun d'eux peut fournir deux métaclés qui seront employées pour l'exploration du corpus qui est maintenant « organisé ». Une **métaclé** peut se définir comme une association préférentielle de mots, qui détermine une partie de l'axe sur lequel on peut retrouver les documents qui lui sont liés. La conception de QNOMIS-II est telle que l'organisation des mots et des documents peut apparaître sous forme de graphes à deux dimensions.

Pour commencer, l'idée est d'identifier par les métaclés, les facteurs les plus pertinents pour la localisation les documents recherchés.

Pour cela, il faut parcourir « manuellement » (on retiendra ce terme par commodité pour désigner une opération qui certes a recours à l'informatique mais qui ne peut s'effectuer sans intervention humaine) les différents tableaux de métaclés pour identifier les plus intéressants, et donc ensuite les graphes où se concentrent les documents pertinents.

Notre expérience montre que deux procédures sont à même de rendre cette recherche plus performante : Le choix d'un mot particulièrement intéressant, avec une fois sa localisation obtenue l'exploration des documents liés aux mots « voisins » ; le choix d'un document spécialement pertinent, et la recherche des documents voisins.

La même procédure peut éventuellement être envisagée en choisissant de rechercher un mot qui est le nom d'un auteur considéré comme une référence dans le domaine considéré.

A l'issue de cette procédure, un **corpus de 143 documents** est constitué grâce à l'AFC.

Par précaution, deux recherches additionnelles et éclairées par les résultats de l'AFC ont été menées, sur un site spécialisé en Médecine du Sport et dans MEDLINE®, pour adjoindre des publications non « puisées » par notre équation logique initiale car postérieures au champ temporel de la recherche, ou non référencées dans les bases de données, ou encore n'ayant pas de résumé. Au total, c'est un corpus de 169 documents ayant précisément trait à l'évaluation de l'activité physique chez le sujet âgé qui constituera la base du travail de recherche en médecine générale qui est à l'origine de cette recherche bibliographique.

3. Analyse de la méthodologie

A posteriori, il est aisé de formuler les critiques de la méthodologie employée dans cette recherche bibliographique. Il apparaît évident, à la lumière de cette première expérience, qu'une seconde recherche procéderait de manière différente, sans doute plus méthodique. Mais c'est précisément grâce à ce premier essai que les améliorations possibles se font jour, aussi est-il intéressant d'évaluer l'influence des différentes étapes de la procédure employée dans ce travail afin de pouvoir envisager des améliorations ultérieures.

[4] ADDAD : Association pour le Développement De l'Analyse de Données.
[5] Conçu par Michel Kerbaol et réalisé par Régis LECLERC, Log INSERM © 2000.

3.1. L'équation initiale

Conçue au départ pour être peu sélective, c'est-à-dire pour éviter autant que possible d'ignorer des articles intéressants, elle prend en compte deux « concepts » qui doivent être associés :

- L'activité physique et ses différentes formulations
- Les domaines intéressant les « sujets âgés »

On peut – toujours *a posteriori* – regretter que la notion d'**évaluation** n'ait pas été incluse dans cette équation, puisque c'est précisément l'objet de la recherche. Cet inconvénient a été compensé dans un second temps par l'adjonction de 1 051 documents complémentaires extraits à l'aide de la notion manquante (mots « questionnaire » et « évaluation »).

3.2. L'élimination des articles sans résumé

Cette élimination répond à la contrainte selon laquelle l'analyse de données ne peut pas porter sur les seuls mots du titre pour être exploitable.

La question de cette élimination reste entière quant à sa légitimation. En effet :

- Le bien-fondé de la sélection d'un article ne peut pas, la plupart du temps, se décider à la lecture du seul titre de celui-ci. De surcroît, le nombre de mots du titre est le plus souvent insuffisant pour permettre une analyse de données textuelles. Les exemples abondent de titres qui paraissent pertinents pour l'objet de notre recherche, ce qui est démenti par la lecture du résumé. Cela pose la question de la rédaction des titres par les auteurs, laissée à leur entière discrétion.
- Certains articles trouvés dans un second temps montrent un probable intérêt pour l'objet de notre recherche, bien qu'ils ne soit pas possible de prendre connaissance d'un résumé. C'est ainsi qu'une dizaine d'articles trouvés grâce à la recherche secondaire dans MEDLINE® ont été inclus dans le corpus des 169 articles retenus.

3.3. Le dédoublement des fiches

Rendu indispensable par le fait que les mêmes publications peuvent être référencées dans les deux bases de données MEDLINE® et SCI®, la suppression des fiches obtenues en double s'est avérée imparfaite pour deux raisons :

- D'abord et surtout les différences minimes qui peuvent exister entre deux articles selon la façon dont ils ont été « saisis » par les indexeurs.
- Ensuite par le fait qu'un même auteur pourra publier le même article dans deux revues différentes, et ne pas souhaiter que cette double publication soit identifiable.

L'inconvénient de ce dédoublage imparfait n'est autre que d'encombrer la base de travail, ce qui lors de la recherche manuelle conduit à un travail fastidieux.

3.4. L'imprécision de la recherche initiale

Résultant en partie de la formulation insuffisante de l'équation de recherche de départ, c'est cette imprécision qui a conduit à la seconde recherche par interrogation sur les mots « questionnaires » et « évaluation ».

Cet inconvénient est indépendant de la méthodologie de recherche bibliographique par l'AFC, dont le principe est justement de pallier à l'inévitable flou initial d'une recherche documentaire.

Ceci nous incite à recommander une recherche en deux temps :

- Recherche initiale sur une équation large, avec extraction de documents « pertinents » ;
- Sélection de mots pertinents au sein de ces documents pour établir un dictionnaire de mots utilisés pour la formulation d'une seconde équation plus performante, qui conduit au recueil d'une base de travail complète et fiable.

4. Analyse des résultats

Pour apprécier globalement l'efficience de la recherche bibliographique par l'AFC, une bonne méthode peut consister à renouveler la même recherche par un moyen plus « conventionnel », et à comparer les résultats obtenus.

Il est indispensable de préciser que cette seconde recherche n'est pas réalisée *de novo*, mais à la lumière (intense) de notre première expérience.

4.1. Méthodologie

Nous avons choisi de mener la recherche au sein des deux bases de données de référence, à savoir MEDLINE® et SCI®.

Pour la recherche au sein de la NLM, les contraintes suivantes sont apportées (en utilisant la fonctionnalité « limits » du moteur de recherche PUBMED®) pour se situer dans la perspective de recherche initiale : recherche sur les mots du titre ou du résumé ; âge supérieur ou égal à 65 ans ; recherche dans MEDLINE® (et non sur les autres bases de données de la NLM) ; recherche restreinte au champ biomédical humain ; date de publication entre 1960 et 1999.

La recherche dans la base SCI® a porté sur les publications parues entre 1993 et 1999.

Une nouvelle équation est écrite pour effectuer la recherche, qui prend en compte la nécessité de recueillir spécifiquement les articles ayant trait à l'évaluation de l'activité physique et à la validité de cette évaluation, sans contrainte d'âge :

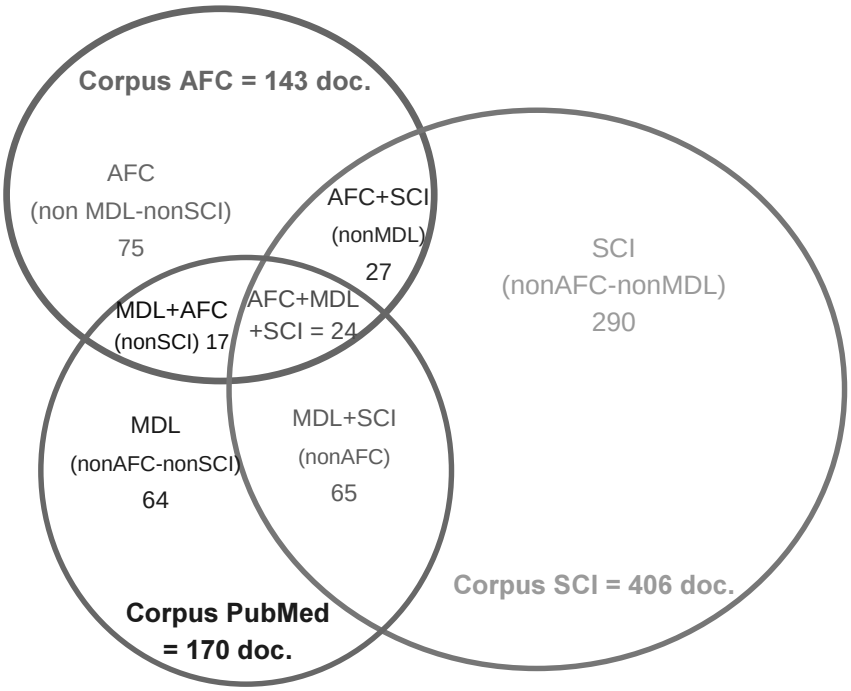
(physical activity)
AND (assess* OR evaluat* OR questionnaire* OR measur*)
AND (reproducib* OR valid* OR reliab* OR repeatab*)

Nous appelons cette équation « équation secondairement optimisée » (**ESO**) pour la différencier de la première.

Cette équation, identique lors des deux recherches comparatives dans les deux bases de données MEDLINE® et SCI®, fournit deux nouveaux corpus bibliographiques : **170** articles extraits de la base Medline®, et **406** articles extraits de la base SCI®.

4.2. Résultats

Le triple croisement entre le corpus bibliographique obtenu par l'AFC (143 documents), celui extrait de Medline (170 documents) et celui extrait de SCI (406 documents) est résumé par le schéma page suivante.



Il convient ensuite d'étudier zone par zone les documents pour évaluer les avantages et les inconvénients de la recherche bibliographique par l'AFC.

On constate d'emblée que **75 articles n'ont été extraits que par l'AFC**, alors qu'ils n'ont été trouvés par l'ESO ni dans Medline®, ni dans SCI®. Les articles trouvés par les trois méthodes ne sont « que » 24. Les articles trouvés conjointement par l'AFC et par l'ESO dans Medline® sont au nombre de 41, ceux trouvés conjointement par l'AFC et par l'ESO dans SCI® sont 51.

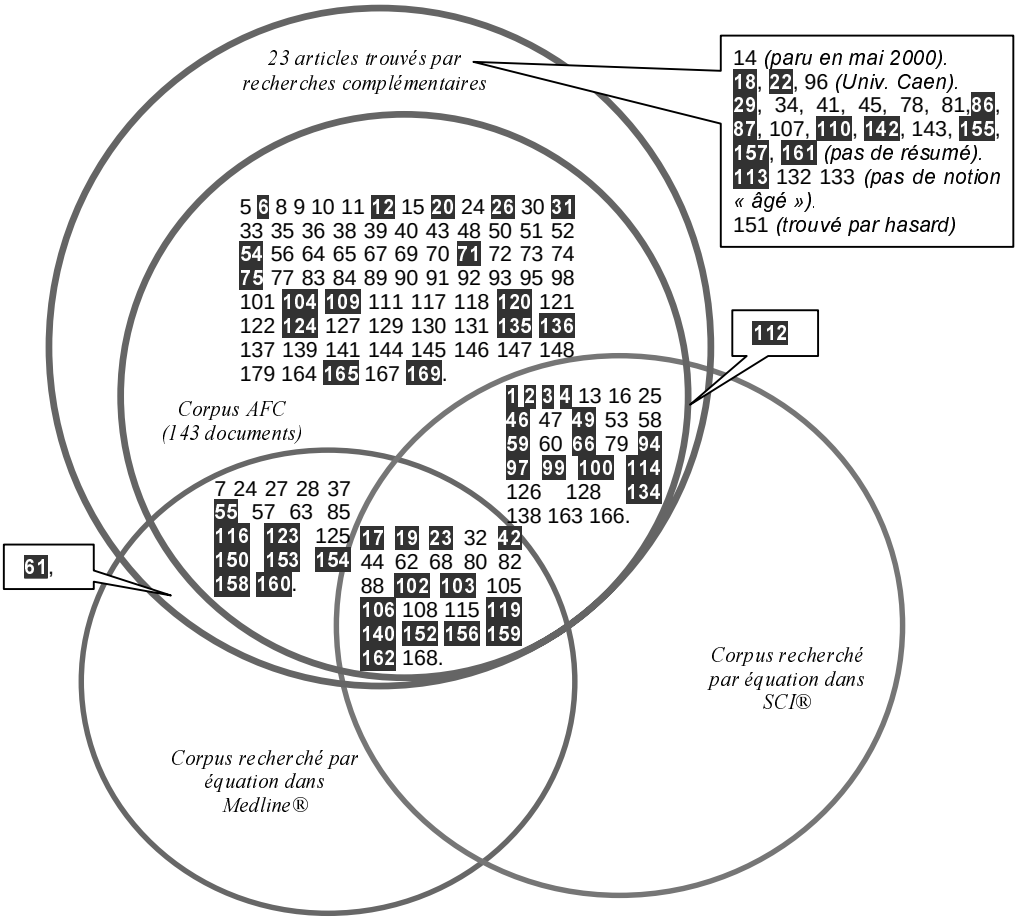
Les questions que l'on doit se poser sont les suivantes :

- Les publications qui n'ont pas été trouvées par l'ESO dans Medline® et/ou dans SCI, et qui ont été extraits de la base de travail par l'AFC sont-ils parmi les plus pertinents pour l'objet de notre recherche ?
- Où peut-on localiser les articles « essentiels » ? En d'autres termes, serait-il utile dans notre recherche bibliographique d'utiliser à la fois l'AFC et la méthode traditionnelle (en gardant en mémoire que dans notre exemple la recherche secondaire « traditionnelle » a été menée à la lumière de la première démarche) pour identifier à coup sûr les articles les plus pertinents ?
- Existe-t-il des articles pertinents dans les corpus obtenus par l'ESO dans Medline® et dans SCI® qui n'auraient pas été identifiés par l'AFC ?
- Dans quelle zone se situe le plus grand « déficit » de recherche bibliographique : dans la zone AFC-SCI(nonMDL), dans la zone AFC-MDL (nonSCI), ou dans la zone MDL-SCI(non AFC) ?
- Existe-t-il des silences inacceptables pour l'ensemble des méthodologies ?

Dans un premier temps, il est possible de répondre aux quatre premières interrogations à condition d'examiner précisément la pertinence de chacun des articles des différents corpus.

Nous prendrons donc en compte prioritairement les 65 articles considérés comme « très intéressants » dans l'exploitation de cette bibliographie, c'est à dire les références premières ou « **articles essentiels** » (AE) dans l'évaluation *validée* de l'activité physique.

On détaille également le corpus bibliographique en faisant figurer dans le graphe ci-après les 26 articles trouvés par la recherche additionnelle, qui complètent les 143 premiers articles trouvés par l'AFC.



Analyse des références principales : les numéros renvoient au classement alphabétique par auteur du corpus bibliographique. Les numéros surlignés représentent les « articles essentiels (AE) », en pratique ceux qu'il est indispensable de ne pas oublier dans un recueil bien conduit.

Parmi les 65 articles « essentiels » (AE) :

- 16 AE sont trouvés exclusivement par la méthode de l'AFC, l'ESO ne les ayant pas extrait de MEDLINE® ni de SCI® ;
- 14 AE sont trouvés à la fois par l'AFC et par l'ESO dans SCI® ;
- 8 AE sont trouvés par l'AFC, et par l'ESO dans MEDLINE®,
- 13 AE sont obtenus également par les 3 méthodes ;
- 1 AE est trouvé par l'ESO dans SCI®, et seulement dans la recherche additionnelle à celle de l'AFC, donc pas par cette dernière ;
- 2 AE sont trouvés par l'ESO dans MEDLINE®, et seulement dans la recherche additionnelle à celle de l'AFC, donc pas par cette dernière ;

– 11 AE ne sont trouvés par aucune des trois méthodes, mais seulement par la recherche additionnelle.

Dans notre recherche, une équation comme l'ESO, avec toutes les réserves précédemment mentionnées au sujet de sa réalisation « éclairée », n'est en mesure de recueillir *au mieux* que 28 articles sur 65 considérés comme *essentiels* (soit 43 %, dans SCI®), voire seulement 21 sur 65 (soit 32 %, dans MEDLINE®). Ces proportions peuvent légitimement être considérées comme très insuffisantes.

4.3. Les silences

Au sein des 26 publications qui ont été ajoutées à celles recueillies par l'AFC pour constituer le corpus étudié (la zone « recherches complémentaires » sur le graphe), il est intéressant d'examiner dans le détail 14 articles « essentiels » (AE) :

– Un AE (n° 112 sur le graphe) est trouvé par l'ESO dans SCI®, et n'a pas été trouvé par l'AFC. Il s'avère que cet article est une *double publication* du n° 113, et que l'absence de la notion « âgé » dans le titre et le résumé, notion exigée dans l'équation initiale de recrutement de notre base de travail, a exclu de cette dernière.

– Un AE (n° 61), trouvé par l'ESO dans MEDLINE®, n'est pas retrouvé par l'AFC. Son exclusion de la base de travail s'explique également par l'absence de la notion « âgé ».

– Un second AE (n° 76), trouvé par l'ESO dans MEDLINE®, n'est pas retrouvé par l'AFC. Il n'existe aucune raison qu'il soit absent de notre base de travail. Seule une erreur dans la manipulation des fichiers informatiques nous semble devoir être évoquée.

Onze AE n'ont été trouvés par *aucune méthodologie*. Il s'agit respectivement :

– de 2 AE (n° 18 et 22) parus dans une revue non référencée par les bases SCI et MEDLINE®, donc par définition ne figurant pas dans la base de travail initiale ;

– de 8 AE (n° 29, 86, 87, 110, 142, 155, 157 et 161) ne possédant pas de résumé, donc qui ne pouvaient être soumis à l'AFC, et qui ont été exclu de notre base de travail initiale ;

– d'un AE (n° 113) exclu de la base de travail pour l'absence de la notion « âgé », raison identique à celle qui a éliminé le n°112 dont il est une double publication.

Une dernière investigation peut être conduite au sujet de publications utilisées dans l'exploitation de la bibliographie, qui ont été considérés comme pertinents par l'Unité de Biologie et de Médecine du Sport du CHU de Rennes. Dans la mesure où leur recueil est totalement indépendant de notre procédure, leur examen permet d'identifier d'éventuels « silences » de l'ensemble de notre méthodologie.

On constate qu'au sein des 15 publications en question :

– 10 articles appartiennent au supplément 9 du numéro 20 (paru en septembre 2000) de la revue *Medicine and Science in Sports Exercise*, donc postérieurement à notre recherche bibliographique ;

– 1 article est paru dans le numéro 6 (juin 2000) de la même revue, donc n'a pu être trouvé pour la même raison ;

– 3 articles ne comportaient pas la notion « âgé » ;

– 1 article, paru en 1997, ne possède pas de résumé.

Reste la question importante des « silences » qui pourraient être détectés à l'examen des corpus recueillis dans SCI® ou dans Medline®, c'est-à-dire dans les zones de publications non trouvées par l'AFC, soit les zones du graphe MDL(nonAFCnonSCI), MDL-SCI(nonAFC) et SCI(non AFCnonMDL). Encore faut-il que si des publications apparaissent intéressantes, elles possèdent un résumé et que la notion « âgé » figure dans le titre ou le résumé, faute de quoi elles sont forcément éliminées d'emblée.

Une scrutation attentive et exhaustive des deux corpus obtenus par l'ESO montre que, lorsqu'il peut apparaître un intérêt à la lecture du titre d'une publication non trouvée par l'AFC (il existe environ une dizaine d'exemples), la notion d'âge ne figure pas dans le résumé, sans dérogation aucune.

A une seule exception près, pour laquelle une erreur dans les manipulations de fichiers semble la plus probable, l'AFC apparaît donc avoir rempli son rôle avec efficacité dans notre recherche bibliographique, en permettant un recueil quasi exhaustif des documents les plus pertinents, dans la limite des contraintes que cette méthodologie impose, notamment en ce qui concerne la présence d'un résumé.

On observe a contrario que les méthodes traditionnelles de recherche par équation logique dans les bases de données équivalentes, même si l'équation en question est « éclairée », sont manifestement insuffisantes, en ignorant *plus de la moitié des références considérées comme primordiales*.

5. Conclusion

Au terme de cette recherche bibliographique appuyée sur l'utilisation de l'analyse de données textuelles, il nous semble opportun d'apporter quelques précisions.

Les performances de la recherche auraient pu être meilleures lors de l'étape initiale de l'écriture de l'équation logique de recrutement de notre base de travail. La restriction portant sur l'âge, argumentée au départ par l'objet de la recherche, s'est avérée ensuite superflue. En effet, l'expérience (et seulement celle-ci) nous a montré par la suite que le fait de contingenter la recherche au domaine traitant spécifiquement des sujets âgés repose sur l'idée intuitive que l'évaluation de l'activité physique chez l'adulte ne peut pas être transposée au sujet âgé. Ce postulat de départ ne repose en réalité sur aucun argument prouvé au début de la procédure de recherche. Il s'est avéré ensuite que certaines publications pertinentes dans le cadre de notre recherche ont été écartées abusivement du fait de cette inexactitude initiale.

Mais, comme nous l'avons dit, cette imprécision n'a été constatée que dans un second temps, à la lumière notamment de l'écriture de l'équation secondaire conçue à l'étape de l'évaluation de la méthode de l'analyse factorielle de correspondance (AFC).

Ceci ne fait qu'ajouter à la nécessité d'une réflexion préalable, pour bien cerner l'objet d'une recherche bibliographique, quelle que soit la méthode utilisée. Nous ajouterons que cette indispensable réflexion ne nous semble pas pouvoir être du seul ressort du chercheur, et que la collaboration d'un spécialiste de la recherche documentaire s'impose, pour pouvoir cerner précisément l'objet d'une investigation, et le vocabulaire nécessaire pour y accéder.

Il existe plusieurs manières de mener une recherche bibliographique, parmi lesquelles on peut individualiser deux grandes méthodes :

– la recherche conduite grâce à la formulation d'une équation logique précise, dont la conception et la manipulation appartiennent, pour être performantes, au spécialiste de la recherche documentaire ;

– la recherche du non-spécialiste de la documentation, menée sur la base d'une interrogation plus ou moins floue, dont la caractéristique essentielle est qu'*elle se fonde sur une représentation de la question propre au chercheur*.

La méthodologie que nous avons expérimentée ici procède d'une autre stratégie, celle d'une analyse statistique du corpus bibliographique par l'analyse factorielle des correspondances (AFC), dont nous résumons ci-après les avantages, les contraintes et les limites.

5.1. Avantage de la méthode de l'AFC pour la recherche bibliographique

Le bénéfice, évident, est montré par la comparaison effectuée dans notre travail entre les performances de la méthode (pourtant imparfaite) utilisant l'AFC et celles de l'équation optimisée dans les mêmes bases de données. Le « manque à gagner » de la méthodologie « classique » dépasse 50 %, autrement dit plus de la moitié des documents pertinents sont passés sous silence. Répétons ici que le plus préoccupant est que ce « silence » est précisément *silencieux*, ce qui est loin de n'être qu'une évidence : le chercheur ayant réuni sa bibliographie par la méthode classique n'aura *aucune conscience* d'ignorer une partie importante des études de ses collègues travaillant sur le même sujet.

On peut dès lors s'interroger sur les conditions réelles de la collaboration entre les équipes de recherche, dans la mesure où elle repose sur des recherches bibliographiques menées au sein des bases de données par des procédures présentant de tels inconvénients.

5.2. Contraintes de la méthode de l'AFC pour la recherche bibliographique

Il s'agit d'une méthode coûteuse en termes de temps, dans la mesure où, si la procédure initiale de recueil et d'analyse est relativement automatisée, l'analyse secondaire nécessite l'intervention humaine, celle décrite comme « manuelle » dans notre exemple de recherche bibliographique. Remarquons au passage que l'intelligence humaine n'est pas près d'être remplacée ici par des dispositifs informatiques [6]. Il faut néanmoins conclure qu'en terme de qualité, les bénéfices sont assez substantiels pour justifier de consacrer le temps et l'énergie nécessaires à l'obtention d'un corpus complet sur le sujet de la recherche en cours.

L'ingénierie nécessaire à l'utilisation de cette méthode ne représente un désavantage que dans la mesure où l'on néglige celle intégrée aux banques de données biomédicales, qui n'apparaît pas lors d'une recherche en ligne.

La nécessité d'une collaboration entre le chercheur et le spécialiste de l'AFC ne peut être considérée comme une contrainte que parce qu'elle suppose un changement dans des habitudes de travail. Là encore, les bénéfices attendus sont des arguments incontournables pour justifier de cet attelage. Nous insisterons sur la nécessaire collaboration qui doit guider la recherche bibliographique dans son ensemble, le chercheur ayant besoin des compétences du spécialiste, et ce dernier ayant besoin de l'expertise du chercheur, au cours des étapes successives de la recherche.

[6] Pinhas N, *Quelle représentation pour les textes dans les bases de données biomédicales*, Ecole Modulad / SFdS-INRIA, Bases de données et statistique, Le Croisic 17-19 novembre 1999.

5.3. Problèmes bibliographiques non résolus par la méthode de l'AFC

5.3.1. Inconvénients liés à la conception des banques de données

Aussi performante qu'elle puisse se montrer dans notre exemple de recherche, l'analyse de données s'applique à des bases dont elle ne peut pas corriger les défauts.

Il en est ainsi :

- du référencement préférentiel des publications anglo-saxonnes, et de leur contribution majoritaire à la littérature mondiale caractéristique des principales banques de données biomédicales ;
- des doubles publications dont les auteurs évitent le repérage en modifiant les textes des fiches documentaires ;
- des silences inhérents au non-référencement de certaines revues dans les principales banques de données, qui nous ont incité à mener des recherches complémentaires.
- de l'inclusion dans la base de travail initiale de publications n'ayant pas trait au domaine biomédical humain, conséquence directe de l'absence de segmentation de la base SCI®. La conséquence en est un « bruit », que l'analyse de données permet toutefois d'éliminer, aussi cet obstacle est-il minime.

Il va de soi que l'intérêt scientifique d'une publication, autrement dit sa validité, est un critère qui n'est évaluable que par le chercheur. On sait qu'il existe des « scories » de ce point de vue dans la littérature publiée, même lorsque les articles sont revus par les « pairs ».

5.3.2. Inconvénients en relation avec la méthodologie

L'inconvénient majeur d'une recherche bibliographique est d'évidence le « silence », responsable de l'ignorance des travaux portant sur le sujet qui nous intéresse.

Nous avons pu constater lors de notre recherche bibliographique :

- des silences relatifs à l'équation de recherche réalisée pour constituer la base de travail. Nous avons déjà envisagé les moyens de réduire cet inconvénient par la réalisation rationnelle de cette équation initiale, au besoin en réalisant un dictionnaire des mots « candidats », à partir d'une première recherche « exploratrice ».
- des silences résultant de l'exclusion des publications qui ne possèdent pas de résumé dans les bases de données, et qui ne sont donc pas accessibles à une analyse statistique. On répétera ici que la présence d'un résumé n'est pas en soi une garantie de l'adéquation d'une publication avec l'objet de la recherche bibliographique lors de sa lecture, même par le spécialiste du sujet considéré.
- des silences inhérents à l'opérateur de la recherche « manuelle », dont l'incidence sera probablement réduite à l'avenir par les améliorations de l'ergonomie du logiciel Q-NOMIS.

Au terme de ce travail, et après avoir parcouru les avenues et les ruelles de la publication scientifique médicale mondiale sur l'évaluation de l'activité physique humaine, nous avons pu constater le fait suivant.

Un nombre considérable de travaux ayant trait aux effets de l'activité physique sur différents paramètres de la santé humaine ont été publiés. La lecture des résumés ou des textes de ces

publications montre que le plus souvent, l'évaluation de l'activité physique est réalisée sur la base de questionnaires ou d'interrogatoires réalisés à l'occasion de l'étude en question.

Compte tenu de ce qui précède, deux interrogations à notre avis essentielles se font jour :

- Quelle est la valeur de ces publications sur le plan épidémiologique, dans la mesure où la validité du paramètre « activité physique » est appréciée par des outils non évalués ?
- Comment peut-on justifier, à ce niveau de la recherche médicale, d'ignorer (*en ignorant cette ignorance*) plus de la moitié des travaux portant sur le même sujet, pour effectuer une recherche et en publier les résultats ?

On pondérera ce qui précède en observant que ce qui peut être observé au travers des publications référencées n'est pas un bilan exhaustif de la recherche mondiale, dans la mesure où tous les travaux ne sont pas nécessairement publiés, ou bien le sont dans des revues de diffusion confidentielle et/ou non référencées dans les grandes banques de données biomédicales. Il s'agit là d'une littérature « grise », par définition inaccessible de l'extérieur d'un domaine considéré.

Ajoutons que le petit nombre d'équipes identifiées dans notre bibliographie qui travaillent sur l'évaluation de l'activité physique peut induire une certaine redondance dans les références citées par ces dernières, et dans la bibliographie globale sur le sujet.

Références

Archimbaud J., *Bibliographie et recherche documentaire en médecine et pharmacie*. Sandoz Ed. 1970.

Bellemans B. *Évaluation de l'activité physique chez le sujet âgé : analyse de la bibliographie et proposition d'un outil de Médecine Générale*. Thèse de Médecine Générale, Université de Rennes I, Décembre 2000.

Benzécri J.-P. et al. (1973). *L'analyse des données*. Dunod éditeur, Paris.

Benzécri J.-P. (1980), *L'analyse des données, Tome 2 : L'analyse des correspondances*. Paris : Bordas. Ed.

Desrichard Y. *Le dédoublonnage des banques de données bibliographiques*. Documentaliste, 1997. 34, 2 : 82-89.

Eveillard Ph., *Les banques de données bibliographiques*, Rev Prat 2000, 50, 16 (suppl).

Kerbaol M., Bansard J.-Y., (1999). *Pratique de l'analyse des données textuelles en bibliographie*. École Modulad SFdS, INRIA, Bases de données et statistique, Dunod Ed., sous presse.

Kerbaol M., Bansard J.-Y., (2000). *Sélection de la bibliographie des maladies rares par la technique du vocabulaire commun minimum*. JADT 2000 vol. 1, M. Rajman & J.-C. Chappelier Ed., EPFL.

Lebart L., Salem A. (1994). *L'analyse des données textuelles*. Dunod Ed., Paris.

Pinhas N., *Quelle représentation pour les textes dans les bases de données biomédicales*, École Modulad / SFdS-INRIA, Bases de données et statistique, Le Croisic 17-19 novembre 1999.

Pochon B., in *Méthodologie documentaire : Introduction à la lecture et à l'écriture de la littérature scientifique*. Gembloux : Faculté universitaire des Sciences agronomiques, 2000.
ISBN 2-87337-004-1. Consulté en ligne le 24/10/2000 à l'adresse :
http://recoda.fsagx.ac.be/edudoc/manuel.htm

Rouillier J., *Évaluation de l'activité physique chez le sujet âgé : Application de l'analyse des données textuelles à la revue bibliographique de la littérature médicale*. Thèse de Médecine Générale, Université de Rennes I, Décembre 2000.