

Travail de Maîtrise en médecine n°4386

Evaluation de l'atrophie cérébrale en IRM. Etude de la corrélation inter-évaluateur d'une échelle visuelle d'évaluation globale de l'atrophie cérébrale

Etudiante : Hélène Fenter

Tuteur : Pr. Phillippe Maeder, médecin chef

Co-tuteurs : Pr. Patric Hagmann, médecin associé
Dr. Vincent Dunet, Médecin associé, MD, BSc

Expert : Meritxell Bach Cuadra

Lausanne, le 3 décembre 2017

Sommaire

Abstract

Table des matières

1. Introduction
2. Méthodologie
 - 2.1. Recrutement des patients
 - 2.2. Pré-traitement séquences IRM
 - 2.3. Développement de *CVRS Lausanne*
 - 2.3.1 Echelles de départ
 - 2.3.2 Modifications apportées
 - 2.3.3 *CVRS Lausanne*
 - 2.4. Méthode de lecture
 - 2.5. Statistiques
3. Résultats
 - 3.1. Ensemble des patients
 - 3.2. Analyse par groupes
4. Discussion
 - 4.1. Comparaison avec la littérature
 - 4.2. Impact des différents facteurs
5. Conclusions
6. Références

1. Introduction

L'imagerie par résonance magnétique est l'examen de choix pour un bilan de démence ou de troubles cognitifs, autant à but diagnostique que pour évaluer la progression de la maladie, écarter d'autres causes de troubles cognitifs, permettant de démontrer une atrophie cérébrale et d'éventuelles lésions vasculaires [1]. L'atrophie corticale est retrouvée chez les patients avec une maladie d'Alzheimer, avec une atteinte prédominante de l'hippocampe, mais est également présente dans d'autres démences (Démence fronto-temporale (DFT), vasculaire, à corps de Lewy), et accompagne aussi la maladie de Parkinson. Une atrophie corticale globale modérée fait partie du vieillissement normal, alors qu'elle est pathologique chez un jeune patient. L'appréciation de l'atrophie corticale est donc un élément important pour une prise en charge optimisée, quelle que soit la raison pour laquelle l'imagerie a été prescrite. Depuis l'apparition de l'IRM, et auparavant avec le scanner X (CT), de nombreux auteurs (description dans méthodologie) ont proposé des moyens de systématiser l'évaluation de cette atrophie, afin de s'assurer de la reproductibilité des méthodes utilisées, ainsi que de leur fiabilité en les confrontant à la clinique.

Aujourd'hui, alors que nous avons à disposition de plus en plus d'outils technologiques permettent une mesure volumétrique précise du cortex, il semble important de se poser à nouveau la question de la pertinence de l'utilisation des échelles visuelles pour l'évaluation de l'atrophie corticale. C'est pourquoi nous avons consacré cette étude à l'investigation de la reproductibilité de l'estimation de l'atrophie corticale par le radiologue.

2. Méthodologie

2.1 Recrutement des patients

les séquences MPRAGE de 103 patients ont été soumises à l'évaluation des experts. Les séquences ont été sélectionnées rétrospectivement dans la base de données du service de radiologie du CHUV, en interrogeant la base de données pour les examens effectués entre janvier 2013 et décembre 2015. Les critères d'inclusion étaient que les patients présentent une démence, ou des troubles cognitifs mineurs ; pour les dates correspondant à celles des examens utilisées dans notre étude, les IRM morphométriques n'étaient effectués que dans ce contexte. Les critères d'exclusion étaient toute autre atteinte que celle liée à une maladie dégénérative, c'est à dire des lésions vasculaire, inflammatoire, ou tumorales ou toute autre lésion focale majeure.

2.2 Pré-traitement

Une fois les patients sélectionnés, nous avons procédé à une anonymisation des séquences IRM grâce à un logiciel interne d'anonymisation, qui élimine toutes les informations liées au patient y compris l'âge et le sexe et attribue un numéro au patient, par lesquels nous nous y sommes depuis référés. Nous avons toutefois conservé un fichier pouvant servir de clé à l'anonymisation si nécessaire, pour pouvoir réaliser des corrélations par tranches d'âge et en fonction du genre.

Par la suite, lorsque les évaluations ont débuté, nous avons dû éliminer les séquences de 9 patients, en raison d'une mauvaise qualité technique de celles-ci, ou encore de lésions focales initialement non détectées, avec finalement un total de 94 patients inclus dans l'étude.

2.3 Développement de « CVRS Lausanne »

2.3.1 Echelles de départ

Afin de pouvoir comparer les évaluations des séquences IRM faites par les experts neuroradiologues, nous avons dû développer un outil d'évaluation objectif et systématique, dans le but d'uniformiser la démarche inter et intra-évaluateur, mais également d'obtenir des résultats numériques qui soient statistiquement comparables entre eux. Pour commencer, nous nous sommes intéressés aux différents outils existant, qui sont utilisés au quotidien, formellement ou informellement, par les neuroradiologues pour évaluer des séquences IRM. Parmi ces échelles, celles auxquelles nous nous sommes particulièrement intéressés sont celles de la « *LADIS study* » [2], *Scheltens hippocampus rating scale*[3], *Scheltens White Matter Changes scale*[4], *O'Donovan's posterior atrophy scale*[5], *Koedam's posterior atrophy scale*[6], *the Fazekas and Schelten's scale*[7], *Victoroff's cortical atrophy scale*[8], and *Pasquier's global atrophy scale*[9]. Le problème auquel nous nous sommes rapidement heurtés, est l'absence d'échelle d'évaluation validée comprenant à la fois toutes les régions du cerveau que nous cherchions à évaluer, c'est-à-dire l'ensemble du cortex. Une des études clés pour cette partie de notre travail a été l'étude coréenne de *Jae-Won Jang et al.* [10], sur laquelle nous nous sommes basés pour le développement de notre propre échelle d'évaluation. Cette étude avait pour but de créer et de valider une échelle « *Comprehensive visual rating scale* » (CVRS) en la confrontant à des résultats de mesures volumétriques, chez des patients avec une cognition normale, un « *mild cognitive impairment* » (MCI) ou encore une maladie d'Alzheimer. Pour ce faire, plusieurs échelles ont été mises ensemble, comprenant l'échelle de Scheltens [3] pour l'évaluation de l'hippocampe, une échelle de Victoroff [8] modifiée pour l'évaluation du cortex, assimilée à une échelle de Koedam [6] pour la partie pariétale, l'évaluation de l'élargissement ventriculaire par O'Donovan [5], l'estimation des hyper-intensités de la matière blanche par l'échelle de Fazekas et Schelten [7] et celle des lacunes et micro saignements grâce à l'échelle de la « *Rotterdam scan study* » [10]. L'ensemble étant noté par points, le maximum de points signifiant le maximum d'atrophie, sur des séquences axiales et coronales pour l'ensemble des structures observées. Les résultats de l'étude démontrent que la CVRS est utilisable en clinique pour évaluer les modifications structurelles du cerveau, avec une bonne congruence inter et intra-observateur. C'est donc sur la base de cette échelle que nous avons basé la « *Comprehensive Visual Rating Scale* » que nous avons par la suite utilisée pour cette étude.

2.3.2 Modifications apportées « CVRS Lausanne »

En utilisant comme base la CVRS de *J.-W. Jang et al.* [10], nous avons donc adapté l'échelle pour qu'elle corresponde aux besoins de l'étude, tout en essayant de ne pas trop s'éloigner du modèle validé. Nous avons donc conservé le support divisé en quatre parties ; la première avec l'échelle de Scheltens [3] modifiée pour évaluer l'atrophie de l'hippocampe, la deuxième avec l'échelle de Victoroff [8] modifiée pour évaluer l'atrophie corticale des lobes frontal et temporal, l'échelle de Koedam [6] modifiée pour l'atrophie pariétale, ce à quoi nous avons rajouté une évaluation de l'atrophie du lobe occipital, avec les mêmes critères

que ceux utilisée par les deux précédentes échelles, c'est-à-dire une observation de l'élargissement des sillons. La troisième partie utilise, comme pour la CVRS initiale, le score de O'Donovan [4] pour l'élargissement ventriculaire, mais au lieu d'évaluer séparément la corne antérieure et la corne postérieure des ventricules, notre échelle sépare le ventricule gauche du ventricule droit. Pour la quatrième partie, nous avons éliminé l'évaluation des lacunes et des micro saignements qui se faisaient sur des séquences FLAIR (séquences que nous n'avons pas utilisées) et n'avons conservé que les hyper-intensités de la matière blanche, dont le score se fait grâce à l'échelle de Fazekas et Scheltens [6], et peut être évalué sur séquences MPRAGE.

Les divergences majeures avec la CVRS de base sont que nous avons décidé, contrairement à *J.-W. Jang et al.* [9], qui proposent l'évaluation en coupes coronales ou axiales pour l'évaluation de l'hippocampe et des différents lobes du cortex, de n'évaluer les séquences qu'en coupes axiales uniquement, dans le but d'uniformiser la procédure et de rendre ainsi les divergences entre les experts plus facilement quantifiables. De plus, le but de l'étude étant de comparer les scores des évaluateurs entre eux, et pas de valider l'échelle utilisée contre la clinique, nous n'avons pas besoin du total des points, contrairement au but initial de ces échelles qui est de donner une estimation numérique de l'atrophie corticale d'un patient.

Pour l'élaboration de notre CVRS, nous nous sommes également penchés sur les recommandations de l'étude de L. Harper [11], qui a mis en évidence l'importance du choix des images de références, ainsi que de la réduction du nombre de points à attribuer dans chaque sous-catégorie scorée, pour améliorer la reproductibilité de l'échelle, mais également de l'effet de l'entraînement des experts.

2.3.3 « CVRS Lausanne »

Voici l'échelle que nous avons assemblée, telle qu'elle a été présentée à nos experts, dans le but de guider leur évaluation, procédant par région, avec, pour chaque zone à évaluer, le score des points à attribuer selon l'estimation de l'atrophie, ainsi qu'une courte phrase indiquant les structures sur lesquelles se concentrer.

1. Atrophie de l'hippocampe

Les éléments à évaluer sont la largeur de la fissure choroïdienne, de la corne temporale et la hauteur de l'hippocampe [Figure 1].

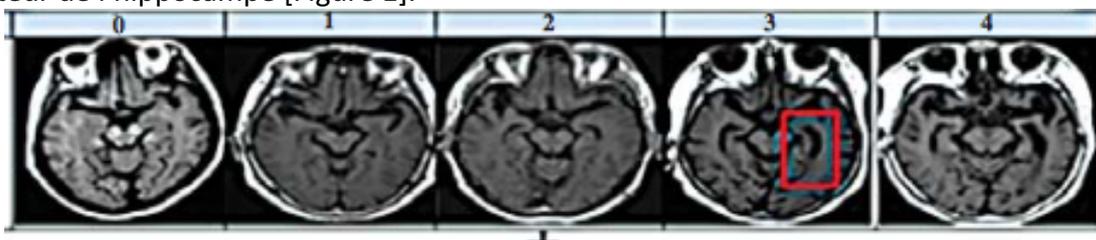


Figure 1 : Image provenant de *Jang J-W et al.* [10], utilisée ici comme référence.

0 : les trois éléments à observer sont normaux

1 : léger élargissement de la fissure choroïdienne, sans atteinte des autres structures

2 : élargissement de la fissure choroïdienne et légèrement de la corne temporale, avec légère diminution de la hauteur de l'hippocampe

3 : fort élargissement de la fissure choroïdienne, avec nette modification de la corne temporale et de l'hippocampe

4 : élargissement majeur de la fissure choroïdienne et de la corne temporale, avec diminution majeure de la hauteur de l'hippocampe

2. Atrophie corticale

Les lobes sont évalués séparément, et le score se base sur l'atrophie des gyri, et la dilatation des sillons, en essayant de ne pas se laisser influencer par la fissure Sylvienne.

Pour chaque lobe, les deux hémisphères sont évalués séparément, par un score de 0 à 3 [Figure2].

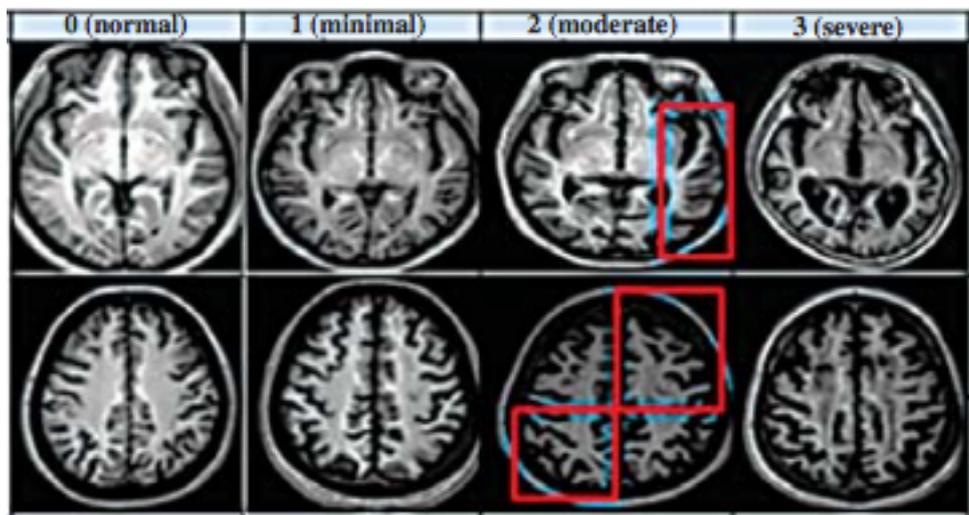


Figure 2 : Image provenant de *Jang J-W et al.* [10], utilisée ici comme référence.

0 : sillons et gyri normaux

1 : léger élargissement des sillons et atrophie des gyri

2 : élargissement et atrophie importants

3 : stade final d'atrophie des gyri du lobe et élargissement majeur des sillons

- A. Lobe frontal
- B. Lobe temporal
- C. Lobe pariétal
- D. Lobe occipital

3. Elargissement des ventricules

La largeur des cornes latérales des ventricules est l'élément à observer, et un score de 0 à 3 points lui est accordée [Figure 3].

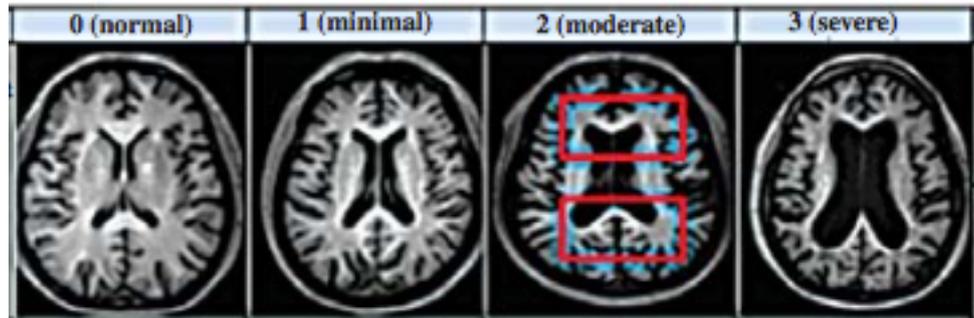


Figure 3 : Image provenant de *Jang J-W et al.* [10], utilisée ici comme référence.

- 0 : pas d'élargissement des cornes latérales des ventricules
- 1 : élargissement mineur des ventricules
- 2 : élargissement significatif des ventricules
- 3 : élargissement majeur des ventricules

4. Hyperintensités de la matière blanche

L'estimation des hyperintensités de la matière blanche est à noter en fonction de la taille de celles-ci.[Figure 4].

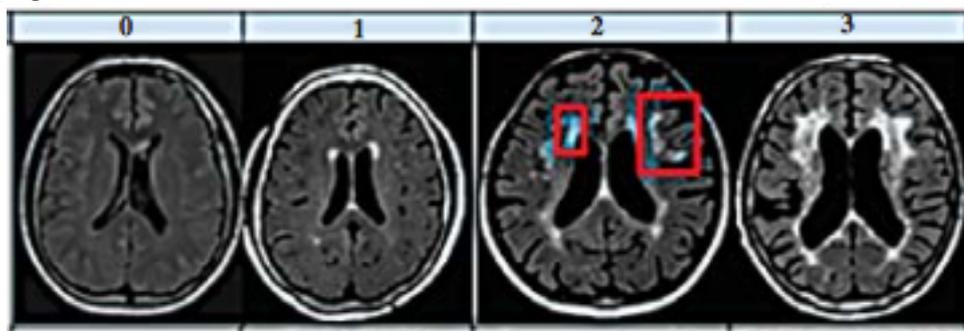


Figure 4 : Image provenant de *Jang J-W et al.* [1], utilisée ici comme référence.

- 0 : pas d'hyperintensités
- 1 : diamètre maximal de moins de 10mm
- 2 : diamètre maximal entre 10 et 25mm
- 3 : diamètre maximal de plus de 25mm

2.4 Méthode de lecture

Une fois l'échelle déterminée, les experts ont procédé à une lecture anonymisée, sans informations relatives à l'âge et au sexe des patients, dans un ordre aléatoire, et ont ainsi noté les séquences des 94 patients selon l'échelle développée ci-dessus. Les trois lecteurs indépendants qui ont effectué cette évaluation ont respectivement 5 ans d'expérience pour E1, 15 ans pour E2, et 35 ans pour E3.

Pour faciliter le processus d'attribution de scores, nous avons créé un formulaire en ligne sur lequel après s'être identifiés, ils pouvaient rapidement cocher les cases correspondant aux scores. Ceci a permis de faciliter la collecte des données en liant le formulaire en ligne avec un fichier Excel, dans lequel les scores se sont inscrits simultanément.

Nous avons également créé un fichier avec les informations du sexe et de l'âge des patients en fonction du numéro d'identification des patients, afin de pouvoir prendre en compte ces paramètres dans l'analyse des résultats, même si ces informations n'étaient pas à la disposition des évaluateurs durant l'évaluation des images.

2.5 Statistiques

2.5.1 Tableaux de contingence

Nous avons commencé par établir des tableaux de contingence [Table 1, 2 et 3] pour les différentes paires d'examineurs. Pour les obtenir, nous avons utilisé le logiciel de statistiques STATA [12], dans lequel nous avons importé notre base de données codée.

Accord : Reflété par les tableaux de contingence l'*accord* entre deux évaluateurs représente la fréquence à laquelle deux observateurs attribuent le même score au même patient, ceci sans prendre en compte la probabilité qu'ils aient donné la même note par hasard.

En parallèle avec les tableaux de contingence, ont donc été déterminés les *accords* entre les évaluateurs, qui sont de 48,9% pour *E1* et *E2*, de 52,2% pour *E1* et *E3*, et de 60,8% pour *E2* et *E3*.

Tableau de contingence E1 - E2

Ex. 1 \ Ex. 2	0	1	2	3	4	Total
0	129	195	31	0	0	355
1	21	238	234	15	0	508
2	1	39	167	50	4	261
3	0	2	17	49	11	79
4	0	0	0	4	15	19
Total	151	474	449	118	30	1222

Table 1 : Tableau de comptabilisation pour les scores attribués par *E1* et *E2*.

Accord E1 et E2 : 48.9%.

Tableau de contingence E1 - E3

Ex. 1 \ Ex. 3	0	1	2	3	4	Total
0	150	170	34	1	0	355
1	34	248	211	15	0	508
2	1	38	168	54	0	261
3	0	1	15	57	6	79
4	0	0	0	4	15	19
Total	185	457	428	131	21	1222

Table 2 : Tableau de comptabilisation pour les scores attribués par *E1* et *E3*.

Accord E1 et E3 : 52.2%.

Tableau de contingence E2 - E3

Ex. 2 \ Ex. 3	0	1	2	3	4	Total
0	105	42	4	0	0	151
1	77	286	104	7	0	474
2	3	117	276	53	0	449
3	0	12	43	59	4	118
4	0	0	1	12	17	30
Total	185	457	428	131	21	1222

Table 3 : Tableau de comptabilisation pour les scores attribués par E2 et E3.
 Accord E2 et E3 : 60.8%.

2.5.2 Coefficients kappa

Puis nous avons recherché les *congruences* inter-observateurs, qui peuvent être calculées grâce au *kappa de Cohen* [13], en pondérant la probabilité qu'un même résultat soit atteint par hasard. Pour calculer les kappas, nous avons utilisé le logiciel STATA [12], qui utilise la méthode de *Fleiss, Levin et Paik* [14], pour les trois paires d'évaluateurs, et également pour les trois évaluateurs combinés en un seul kappa.

Formule du kappa de *Fleiss, Levin et Paik* :

$$\hat{\kappa} = \frac{B - W}{B + (\bar{m} - 1)W}$$

A noter que selon *Landis et Koch* [15], il est établi qu'un coefficient kappa de <0.4 est considéré comme une mauvaise *congruence*, un kappa de 0.4 à 0.8 est considéré comme une *congruence* modérée, un kappa de >0.8 comme bonne *congruence*, et les kappas négatifs sont significatifs d'une moins bonne congruence que le hasard pourrait laisser supposer.

Les kappas ont donc été calculés en fonction de différents facteurs que nous avons choisi de mettre en évidence, afin de déterminer leur éventuelle influence sur la reproductibilité de l'échelle. Ces facteurs ainsi isolés sont les suivants : les différentes paires d'évaluateurs, pour mettre en évidence une éventuelle courbe d'apprentissage, les différentes régions du cortex évaluées, le genre des patients et leur âge.

3. Résultats

3.1 Kappa tous scores

Les kappas ont été calculés tout d'abord pour les trois évaluateurs [Table 4], pour l'ensemble des patients, en fonction d'une des 13 régions examinées. Le kappa global pour l'ensemble des régions est de 0.35. Ceci correspond à un relativement mauvais *inter-rater agreement*.

Region	<i>kappa</i>		Whole Brain
	Right	Left	
Hippocampus	0.42	0.41	
Frontal	0.35	0.32	
Temporal	0.12	0.12	
Parietal	0.40	0.33	
Occipital	0.21	0.22	
Ventricle Dilatio	0.39	0.37	
WMH			0.39

Table 4 : Kappas de l'ensemble des scores.

Kappas par paires

Nous avons ensuite déterminé les kappas par paires [Tables 5, 6 et 7], en procédant par région, comme précédemment. Les examinateurs sont *E1*, *E2* et *E3*. Les kappas globaux [Table 8], c'est-à-dire prenant en compte l'ensemble des régions pour tous les patients évalués par les observateurs considérés, ont également été calculés.

Region	<i>kappa E1 et E2</i>		Whole Brain
	Right	Left	
Hippocampus	0.38	0.40	
Frontal	0.24	0.18	
Temporal	0.16	0.21	
Parietal	0.27	0.20	
Occipital	0.17	0.17	
Ventricle Dilation	0.25	0.26	
WMH			0.45

Table 5 : Kappas des examinateur *E1* et *E2*.

Region	<i>kappa E1 et E3</i>		Whole Brain
	Right	Left	
Hippocampus	0.46	0.44	
Frontal	0.36	0.32	
Temporal	0.08	0.04	
Parietal	0.41	0.43	
Occipital	0.16	0.17	
Ventricle Dilation	0.45	0.44	
WMH			0.34

Table 6 : Kappas des examinateurs *E1* et *E3*.

Region	kappa E2 et E3		
	Right	Left	Whole Brain
Hippocampus	0.42	0.40	
Frontal	0.47	0.49	
Temporal	0.22	0.20	
Parietal	0.53	0.38	
Occipital	0.40	0.42	
Ventricle Dilatation	0.50	0.45	
WMH			0.39

Table 7 : Kappas des examinateurs E2 et E3.

kappa	Groupes			
	E1 / E2	E2 / E3	E1 / E3	E1 / E2 / E3
	0.29	0.44	0.33	0.35

Table 8 : Kappas globaux par groupes d'examineurs.

Nous avons ainsi remarqué que les examinateurs E2 et E3 avaient la meilleure concordance, avec un kappa global de 0.44, même si celui-ci correspond à un *inter-rater agreement* modéré, il est meilleur que les deux autres paires. Cette observation est validée par l'*accord* (ou *complete agreement*) de 60.8% entre E2 et E3, déterminé grâce au tableau de comptabilisation [Table 3]. La plus mauvaise *congruence* est celle entre E1 et E2, avec un kappa global de 0.29, avec un *accord* de 48.9% [Table 1], signifiant une mauvaise reproductibilité entre ces évaluateurs. Ces différences pourraient être expliquées par une courbe d'apprentissage, E1 étant le plus jeune des experts, ceci pourrait expliquer le moins bon *agreement* avec E2 qui est l'expert le plus expérimenté, et ainsi entre les résultats de E2 et E3.

Kappas par région

Dès le départ, les kappas ont été calculés par région observée [Table 4]. Nous avons donc pu observer des variations en fonction de la région ; celles de l'hippocampe (kappa 0.41 et 0.42) et du lobe frontal (kappas 0.35 et 0.32) généraient une meilleure reproductibilité, alors que celle-ci était moins bonne pour les régions temporales (kappas 0.12 et 0.12) et occipitales (kappas 0.21 et 0.22).

Une autre observation qui a pu être faite en observant les différentes régions, est que pour 10/13 d'entre elles, les scores extrêmes, c'est-à-dire ceux correspondant à une atrophie inexistante ou très faible (score de 0 ou 1), ou ceux correspondant à une atrophie importante (maximal de 3 ou 4 selon la région observée), étaient source de kappas plus élevés, alors que ceux relatifs à une atrophie modérée généraient des kappas plus faibles.

<i>Right Hippocampus</i>		<i>Left Hippocampus</i>	
score	kappa	score	kappa
0	0.66	0	0.70
1	0.36	1	0.42
2	0.19	2	0.29
3	0.32	3	0.32
4	0.73	4	0.49
combined	0.42		0.41

Table 9 : Kappas de la région de l'hippocampe, démontrant une meilleure reproductibilité aux extrêmes.

Ce phénomène est ici illustré par les kappas des deux régions hippocampiques [Table 9], mais peut également être observé dans les autres régions, excepté pour le lobe temporal, des deux côtés, et la partie gauche du lobe occipital.

3.2 Kappas par groupes

Afin de pouvoir observer uniquement l'effet que les différents facteurs peuvent avoir sur les corrélations entre évaluateurs, nous avons décidé d'utiliser uniquement les données des examinateurs E2 et E3 pour la partie qui suit. Les facteurs que nous avons choisi d'isoler sont le genre du patient [Tables 10 et 11] et l'âge du patient [Tables 12,13,14 et 15]

Kappas par sexes

Region	<i>kappa (E2 - E3) pour les hommes</i>		
	Right	Left	Whole Brain
Hippocampus	0.41	0.36	
Frontal	0.57	0.52	
Temporal	0.17	0.28	
Parietal	0.46	0.28	
Occipital	0.31	0.39	
Ventricle Dilatio	0.51	0.47	
WMH			0.33

Table 10 : Kappas des patients de sexe masculin.

kappa (E2 - E3) pour les femmes

Region	Right	Left	Whole Brain
Hippocampus	0.44	0.43	
Frontal	0.40	0.46	
Temporal	0.25	0.14	
Parietal	0.57	0.44	
Occipital	0.48	0.43	
Ventriple Dilatio	0.47	0.42	
WMH			0.42

Table 11 : Kappas des patients de sexe féminin.

Les kappas sont équivalents s'ils sont séparés par sexe, avec un kappa global pour E2 et E3 de 0.45 pour les kappas des femmes et de 0.42 pour les hommes. Le genre des patients ne peut donc pas être considéré comme un facteur influençant la reproductibilité de l'échelle dans cette étude.

Kappa par âge

Les kappas selon l'âge ont été calculés par quartiles de la distribution selon l'âge ; le premier comprenant les âges en dessous de 67 ans [Table 12], le deuxième entre 67 et 77 ans [Table 13], le troisième entre 77 et 85 ans [Table 14] et le dernier comprenant les patients de plus de 85 ans [Table 15]. Les âges définissant la limite des quartiles sont compris dans les groupes, aussi bien vers le haut que vers le bas.

kappa (E2 - E3) pour les < 67 ans

Region	Right	Left	Whole Brain
Hippocampus	0.58	0.48	
Frontal	0.34	0.34	
Temporal	0.34	0.24	
Parietal	0.40	0.31	
Occipital	0.40	0.40	
Ventriple Dilatio	0.19	0.20	
WMH			0.45

Table 12 : Kappas pour E2 et E3 des patients de <67 ans.

kappa (E2 - E3) pour les 67 - 77 ans

Region	Right	Left	Whole Brain
Hippocampus	0.29	0.37	
Frontal	0.54	0.46	
Temporal	0.28	0.23	
Parietal	0.48	0.28	
Occipital	0.15	0.29	
Ventriple Dilatio	0.57	0.42	
WMH			0.48

Table 13 : Kappa pour E2 et E3 des patients de 67 à 77 ans.

Region	<i>kappa (E2 - E3) pour les 77 - 85 ans</i>		
	Right	Left	Whole Brain
Hippocampus	0.42	0.27	
Frontal	0.63	0.71	
Temporal	-0.04	0.01	
Parietal	0.62	0.47	
Occipital	0.46	0.34	
Ventriple Dilatio	0.41	0.47	
WMH			0.03

Table 14 : Kappa pour E2 et E3 des patients de 77 à 85 ans.

Region	<i>kappa (E2 - E3) pour les > 85 ans</i>		
	Right	Left	Whole Brain
Hippocampus	0.25	0.32	
Frontal	0.39	0.47	
Temporal	-0.08	-0.01	
Parietal	0.51	0.30	
Occipital	0.46	0.53	
Ventriple Dilatio	0.50	0.44	
WMH			0.36

Table 15 : Kappas pour E2 et E3 des patients de <85 ans.

Le kappa global des patients de moins de 67 ans est de 0.38, celui des patients de 67 à 77 ans est de 0.40, celui des patients de 77 à 85 ans est de 0.41 et celui des patients de plus de 85 ans est de 0.42. Nous observons ainsi une légère amélioration des kappas avec l'augmentation de l'âge des patients.

4. Discussion

4.1 Facteurs influençant les résultats

Les résultats obtenus nous démontrent une reproductibilité inter-évaluateur globale médiocre, la reproductibilité intra-évaluateur n'étant pas testée. Des variations sont observées selon la région du cerveau prise en compte ; la zone de l'hippocampe étant notée de manière plus uniforme que les autres régions, avec un coefficient kappa global (comprenant les scores de l'hippocampe gauche et droit) de 0.42 pour les trois évaluateurs. Cet effet pourrait s'expliquer par le fait que l'hippocampe est une région clé pour l'évaluation de l'atrophie corticale et en conséquent, est plus souvent examinée avec systématique par les neuroradiologues. De plus, l'échelle de Schelten est validée et utilisée depuis longtemps. La région du lobe temporal, en revanche, ainsi que le lobe occipital dans une moindre mesure, avec des kappas globaux relatifs de 0.12 (temporal) et 0.21 (occipital), sont évalués différemment selon les experts. Le lobe occipital est moins systématiquement évalué dans une lecture de routine des séquences IRM, en plus du peu d'échelles disponibles pour le faire. Notre CVRS utilise les mêmes critères que pour les autres lobes, mais

l'élargissement des sillons est peut-être moins flagrant ou plus tardif que pour les autres lobes du cortex.

Pour le lobe temporal, l'explication pourrait être liée à un autre phénomène qui est apparu : les experts sont plus en accord pour les scores extrêmes que pour les scores moyens, et le lobe temporal est, dans notre étude, majoritairement noté avec des scores moyens. Une autre explication pourrait être que l'évaluation de l'atrophie du lobe temporal se fait en routine selon l'échelle de Scheltens, qui évalue simultanément l'atrophie de l'hippocampe et que de noter l'atrophie du lobe temporal en omettant l'hippocampe s'avèrerait peut-être non seulement inhabituel mais aussi de moindre utilité.

Il est également apparu que parmi nos trois experts, deux avaient une faible *congruence*, alors que le troisième se situait entre les deux autres. Une hypothèse que nous avons formulée par rapport à ce phénomène, est que les deux experts avec le plus faible accord, sont le moins expérimenté et le plus expérimenté des neuroradiologues, ce qui laisse envisager une courbe d'apprentissage, sur laquelle le troisième examinateur se situerait également entre les deux autres. C'est pourquoi nous avons utilisé la paire avec la meilleure *congruence* pour l'observation des autres facteurs pouvant influencer les résultats.

La segmentation par âge et par sexe n'a pas amené de différences majeures avec le reste des kappas, mais la progression de l'âge des groupes observés est liée avec une augmentation des kappas de ces groupes. Cette amélioration de l'accord entre les neuroradiologues pour les âges plus élevés semble cohérente selon l'observation que les scores extrêmes correspondent à une meilleure reproductibilité inter-évaluateur : les patients plus âgés ont des scores plus élevés.

4.2 Comparaison avec la littérature

Les résultats de notre étude s'alignent avec ceux de certaines études précédentes, notamment celle publiée en 1997 par P. Scheltens et al [16], qui obtenaient des coefficients kappa globaux de 0.34 et 0.24, ce qui correspond à nos résultats. Toutefois cette étude a utilisé des zones à évaluer beaucoup plus larges, qui comprenaient chaque fois les deux hémisphères, et devaient être évaluées selon la dilatation des sillons, et l'élargissement ventriculaire séparément. Ils ont obtenu des kappas supérieurs (0.39) pour la région frontale, tout comme nos résultats, et plus faibles pour la région temporale (0.29) et occipitale (0.32). Ils ont aussi, et comme beaucoup d'autres études, [11], [2], [6] et [8] une congruence inter-évaluateur supérieure à la congruence intra-évaluateur, avec un kappa de 0.58, démontrant ainsi une bonne reproductibilité intra-évaluateur pour l'utilisation des échelles visuelles. Un autre facteur soulevé, est l'importance d'une explication du fonctionnement de l'échelle juste avant l'étape de notation et non pas à distance. L'étude CERAD, de Davis et al. [17], rapporte des kappas inter évaluateurs de 0.64 à 0.82 mais aussi un moins bon accord entre les évaluateurs au niveau du lobe temporal, phénomène que nous avons également mis en évidence dans notre étude, tout comme P. Scheltens et al. [16]. L'explication développée était que les images données comme exemples n'étaient pas dans le même plan que les séquences à évaluer, mais également que les experts avaient une manière d'appliquer les échelles, relevant d'habitudes personnelles, qui était systématiquement légèrement différente.

L'étude de Victoroff et al [8], a obtenu des kappas de 0,62 pour la région frontale, 0.38 et 0.2 pour la région temporale, et 0.54 pour la région pariétale, suivant le même schéma de

meilleure reproductibilité pour le lobe frontal et moindre pour le lobe temporal. De plus ils ont pu démontrer une courbe d'apprentissage importante, en comparant une paire d'examineurs ayant contribué à l'élaboration de l'échelle entre eux, et ceux-ci ont obtenu des kappas de 0.83 (frontal), 1.0 et 0.64 (temporal), et 0.8 (pariétal). Les postulations suivantes ont été émises dans le rapport : la comparaison d'images de cortex (comparaisons avec images de référence) pourrait être difficile à effectuer pour l'œil humain en raison des circonvolutions, l'utilisation de multiples images à évaluer ou simplement de ne pas fixer une coupe standard à évaluer selon la structure, pourrait rendre le processus plus difficile et par conséquent moins reproductible. Pour finir, la question de différentes acquisition d'imagerie (différentes machines ou centres) a été jugée source probable de discordance supplémentaire.

Plusieurs autres études ont retenu une meilleure reproductibilité que la notre, notamment celle de Koedam et al [6] avec un kappa inter-évaluateur global de 0.73, et des kappas de 0.91 pour le lobe médio-temporal et de 0.7 pour l'atrophie postérieure, pour une évaluation qui a été faite avec l'information de l'âge et du genre du patient. Une autre étude de « LADIS study » [19] a obtenu un kappa inter-évaluateur de 0.7 pour l'atrophie corticale globale. O'Donovan et al. [5] ont eu des résultats également plus positifs ; un kappa inter-évaluateurs de 0.9 pour l'évaluation de l'atrophie médio-temporale, 0.85 pour la région postérieure et 0.9 pour l'élargissement des ventricules., toutefois ces coefficients ont été calculés pour deux évaluateurs dont un avait déjà évalué les mêmes images en connaissances des informations des patients et de leur diagnostic.

Pour l'évaluation des hyperintensités de la matière blanche, une autre étude de P. Scheltens et al [4] a jugé que l'échelle de Fazekas [6] (que nous avons utilisée dans un version modifiée pour l'évaluation de la matière blanche) est une de celles offrant la meilleure reproductibilité, avec toutefois des kappa de 0,43 pour la matière blanche péri-ventriculaire et de 0.5 pour la matière blanche sous-corticale. La « Rotterdam Scan Study » [18] qui a été conduite pour étudier les lésions de la matière blanche a obtenu des kappas de 0.79 pour la matière blanche péri-ventriculaire, et 0.88 pour la matière blanche sous-corticale. Comme notre étude, celles de I. Harper et al [11] et de Victoroff et al. [8] ont mis en évidence une amélioration de la reproductibilité avec l'entraînement des observateurs à utiliser les échelles.

D'autres études, comme celle de la « LADIS study » [2] et la la CVRS de J.W Jang [10], ont procédé à une confrontation avec une analyse volumétrique, les deux obtenant des résultats soutenant une équivalence des échelles visuelles à la volumétrie, le kappa global de l'étude de J.W Jang étant de 0.94, et supérieur à leurs résultats de volumétrie, alors que pour la « LADIS study », deux échelles de complexité différente étaient aussi performantes que la volumétrie lorsque confrontées à la clinique.

5. Conclusion

Au vu de ces résultats, il apparaît clairement que les échelles visuelles comme outil de diagnostic de l'atrophie corticale sont bien reproductibles pour noter l'absence d'atrophie, ou une atrophie massive, mais médiocres pour discriminer les degrés d'atrophie moyens, qui sont pourtant d'une haute importance clinique, autant pour diagnostiquer une démence débutante que pour suivre l'évolution d'une atteinte connue. Ces échelles sont

désormais à reconsidérer, maintenant que d'autres options automatisées sont en train de devenir disponibles, avec dans le futur une place croissante pour ces nouvelles technologies.

6. Références

1. Fazekas F, Chawluk JB, Alavi A, Hurtig HI, Zimmerman RA. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *AJR Am J Roentgenol.* 1987 Aug;149(2):351–6.
2. Gouw AA, Van der Flier WM, van Straaten ECW, Barkhof F, Ferro JM, Baezner H, et al. Simple versus complex assessment of white matter hyperintensities in relation to physical performance and cognition: the LADIS study. *J Neurol.* 2006 Sep;253(9):1189–96.
3. Scheltens P, Leys D, Barkhof F, Huglo D, Weinstein HC, Vermersch P, et al. Atrophy of medial temporal lobes on MRI in “probable” Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. *J Neurol Neurosurg Psychiatry.* 1992 Oct;55(10):967–72.
4. Scheltens P, Erkinjuntti T, Leys D, Wahlund LO, Inzitari D, del Ser T, et al. White matter changes on CT and MRI: an overview of visual rating scales. *European Task Force on Age-Related White Matter Changes. Eur Neurol.* 1998;39(2):80–9.
5. O'Donovan J, Watson R, Colloby SJ, Firbank MJ, Burton EJ, Barber R, et al. Does posterior cortical atrophy on MRI discriminate between Alzheimer's disease, dementia with Lewy bodies, and normal aging? *Int Psychogeriatr.* 2013 Jan;25(1):111–9.
6. Koedam ELGE, Lehmann M, van der Flier WM, Scheltens P, Pijnenburg YAL, Fox N, et al. Visual assessment of posterior atrophy development of a MRI rating scale. *Eur Radiol.* 2011 Dec;21(12):2618–25.
7. den Heijer T, van der Lijn F, Koudstaal PJ, Hofman A et al. A 10 year follow-up of hippocampal volume on magnetic resonance imaging in early dementia and cognitive decline. *Brain* 133, 163-1172.
8. Victoroff J, Mack WJ, Grafton ST, Schreiber SS, Chui HC. A method to improve interrater reliability of visual inspection of brain MRI scans in dementia. *Neurology.* 1994 Dec;44(12):2267–76.
9. Pasquier F, Leys D, Weerts JG, et al. Inter and intraobserver reproducibility of cerebral atrophy assessment on MRI scans with hemispheric infarcts. *Eur Neurol.* 1996; 36:268-272.
10. Jang J-W, Park SY, Park YH, Baek MJ, Lim J-S, Youn YC, et al. A comprehensive visual rating scale of brain magnetic resonance imaging: application in elderly subjects with

Alzheimer's disease, mild cognitive impairment, and normal cognition. *J Alzheimers Dis.* 2015;44(3):1023–34.

11. Harper L, Barkhof F, Fox NC, Schott JM. Using visual rating to diagnose dementia: a critical evaluation of MRI atrophy scales. *J Neurol Neurosurg Psychiatry.* 2015 Apr 14;jnnp-2014-310090.
12. STATA, obtenu de www.stata.com
13. Cohen J, et al. A coefficient of agreement for nominal scales. *Educational et Psychological Measurement.* 1960. 20:37-46.
14. Fleiss JL et al. Assessing the accuracy of multivariate informations. *Journal of the American Statistical Association.* 1966. 61:403-412
15. Landis J, Koch G. The measurement of observer agreement for categorical data. *Biometrics.* 1977 Mar;33 :159-174
16. Scheltens P, Pasquier F, Weerts JGE, Barkhof F, Leys D. Qualitative Assessment of Cerebral Atrophy on MRI: Inter- and Intra-Observer Reproducibility in Dementia and Normal Aging. *ENE.* 1997;37(2):95–9.
17. Davis P.C, Gray L, Albert M, et al. The consortium to establish a registry for Alzheimer's disease (CERAD). Part III. Reliability of a standardized MRI evaluation of Alzheimer's disease. *Neurology* 1992;42:1676-1680
18. de Groot JC, de Leeuw FE, Oudkerk M, van Gijn J, Hofman A, Jolles J, et al. Cerebral white matter lesions and cognitive function: the Rotterdam Scan Study. *Ann Neurol.* 2000 Feb;47(2):145–51.
19. Jokinen H, Lipsanen J, Schmidt R, Fazekas F, Gouw AA, Flier WM van der, et al. Brain atrophy accelerates cognitive decline in cerebral small vessel disease The LADIS study. *Neurology.* 2012 May 29;78(22):1785–92.